Hybrid Image Retrieval Model Using ResNet-50 and Vision Transformer for Enhanced Feature Representation: A Novel Approach

Moloy Dhar[1]*, Soumyajit Nandi[1], Soham Bhattacherjee[1], Soumyanil Dey[1], Sayantan Dutta[1], Simran Ray[1]

[1] *Department of Computer Science & Engg., Guru Nanak Institute of Technology, Sodepur, Kolkata, India*

***Abstract*— In this paper, we present a composite image retrieval framework that boasts the advantages of both ResNet-50 and Vision Transformers (ViT) to improve image feature extraction and retrieval performance. Conventional convolutional neural network (CNN), such as ResNet-50, is well domain-1 at modeling hierarchical spatial features in images and Vision Transformers are better suited for handling long-range dependencies and Attention-based feature selection. Our hybrid framework integrates ResNet-50 as a feature extractor along with a Transformer-based attention mechanism to facilitate enhanced Embedding of image features. We benchmark our experiments demonstrating improved retrieval performance as compared to ResNet-50 and independent ViT models as well. Our experimental findings show that the hybrid model achieves superior retrieval efficiency especially for more difficult fine-grained image search tasks where similar looking images needs to be separated. We further investigate computational efficiency, Feature separability, and retrieval efficient precision by using principal component analysis (PCA) and various performance measures.**

As part of a new approach, we created custom data records from Westbengalen Craft and GI-marked products by collecting and curating high quality images and handling the lack of data records published in this niche domain. We propose a hybrid image retrieval architecture that can scale and adapt for different image retrieval applications such as visual search engines to medical image analysis.

***Keywords*— CBIR, ResNet-50, ViT, CNN, Patch Embedding, Positional Encoding**

1. INTRODUCTION

With the advent of rapid digital transformation, a variety of applications within areas such as e-commerce, content retrieval, medical diagnostics, and surveillance now routinely leverage image search as an underlying technology. Traditional image search methods rely primarily on handcrafted features, which often do not sufficiently characterize more complex patterns and semantic relationships present in images. As such, approaches based on deep learning, and CNN in particular, have emerged as suitable and powerful solutions for both feature extraction and matching of similarity.

Finally, to streamline the product discovery process, image-based search will become a necessary new feature and paradigm for modern e-commerce applications. Instead of wrestling with browsing through number of categories, users and shoppers will simply need to upload an image or scan a product to quickly and effectively discover items visually similar (or related) items.In this work, we propose a hybrid image search model (Fusion Model) that combines ResNet-50, which is a widely used deep CNN, with Vision Transformer (ViT), in order to take advantage of both

architectures independently. While ResNet-50 provides strong representation learning of low-level spatial feature and local textures, vision transformer learns strong long-range dependencies and global contextual relationships. By combining these two different representations of features we expect to have improved retrieval accuracy, robustness, and efficiency.

We investigate different fusion techniques for combinations of ResNet-50 and ViT features including concatenation, attention-based fusion, and feature aggregation models. The experimental evaluation of benchmark dataset demonstrates that our hybrid model significantly improves retrieval performance compared to independent CNN or Transformer approaches.This study provides a detailed study of our fusion methodology, experimental setup and results, and highlights the advantages of hybrid feature extraction for image searching applications.

2. Literature Survey:

The field of image retrieval in the domain of computer vision has become prominent in paper with multiple methods offered over the years, from conventional handcrafted feature extraction methods to deep learning (DL) -based methods. As such, we will look at related developments within image retrieval methods using CNNs, and ViTs, as well as hybrid models, in this section.

Early image retrieval systems were based on handcrafted features (SIFT [1], HOG [2], and color-histograms [3]). These methodologies performed well in a clean environment, but struggled when facing variations in lighting, scale, and viewpoint. The recent development of DL caused a major increase in retrieval performance by learning deep, high level, and robust features.

The exploration of retrieval-based methods using DL systems drew influence from Convolutional Neural Networks (CNNs) methods. [4] demonstrated the efficiency of CNNs and presented AlexNet which achieved superior performance in comparison to conventional methods for image classification applications. [5] broadened the CNN method with the introduction of ResNet-50 which introduced residual learning allowing the sub-models in the deeper models to have improved learning capability. VGG-16, ResNet-50, and DenseNet all became well-cited CNNs for many applications once CNNs were established within image retrieval tasks [6]. While CNNs allow spatial hierarchies to be maintained within images, CNNs do not have the capability to model dependencies at long distances or contextual relationships which are characteristics of human level performance under increasingly complex image retrieval tasks [7, 8].

Transformers were originally introduced for Natural Language Processing (NLP) by the works of [9], before being re-evaluated in a computer vision setting. Vision Transformers (ViTs) transformed the landscape of image analysis by treating images as sequences of patches and leveraging self-attention mechanisms that allow modeling long-range dependencies between the components (patches) [10]. While CNNs typically only attend to certain portions of the image, ViTs have the potential to attend to the most informative parts of the image, resulting in higher retrieval performance fine-grained datasets [11, 12]. However, ViTs are typically only productive for large-scale datasets and can be computationally expensive compared to CNNs, making them less efficient for applications that require real-time processing [13].

Recent works have explored hybrid models, where CNN extracts features from the input and the output of the network employs Transformer-based separated attention mechanism to make use of the advantages of both architectures. Works like BoTNet [14, 15] and ConViT [16, 17] have

considered combining CNN with attention layers to improve on the representation of features. Furthermore, hybrid models have been notably successful for tasks concerning fine-grained image retrieval and relationship understanding in a visual context [18, 19, 20]. This is similar for our proposed hybrid model, using ResNet-50 as a hierarchical feature extractor and Transformers for spatial attention, which selects for the best mixed-BIT features advancing retrieval performance with a cost on accuracy and efficiency.

3. Methodology

In this paper, we focuses on developing an advanced image retrieval system by leveraging DL models for feature extraction and similarity matching [21]. The methodology consists of multiple key components, including dataset preparation, feature extraction, similarity computation, and performance evaluation. The methodology is outlined as follows:

- Data Collection & Preparation:-

The image dataset is stored in a cloud-based environment (Google Drive) to ensure scalability and accessibility. The dataset comprises images in various formats, including JPG and PNG. Preprocessing steps include:

- Resizing images to a fixed dimension for consistency.
- Normalization to scale pixel values between 0 and 1.
- Data Augmentation (if applicable) to enhance model generalization.

- Machine Learning Models :-

- Resnet 50 Architecture:

ResNet-50 is a deep convolutional neural network that incorporates residual connections to enable the effective training of very deep models [22, 23]. For this paper, we used a pre-trained ResNet-50 model, originally trained on the ImageNet dataset, and adapted it for the identification and classification of handcrafted and artisan items from West Bengal. Its ability to extract detailed hierarchical features proved beneficial in distinguishing complex textures and patterns in the images. This model is widely recognized for its stable performance and relatively lower computational demands compared to newer architectures. By leveraging transfer learning, we minimized training time while achieving high accuracy even with a limited dataset.

- Vision Transformer (ViT):

ViT model, inspired by transformer networks in NLP, brings a novel approach to visual tasks by representing images as sequences of patches. Instead of processing the entire image as a whole, it divides it into fixed-size segments, flattens them, and projects them into a lower-dimensional space [24]. Positional encodings are added to preserve spatial context, and the sequence is then processed by a transformer encoder. This structure allows the model to capture long-range dependencies and complex patterns effectively. After initial training on large-scale image datasets, ViT can be fine-tuned for specific vision applications like classification or retrieval [25].

- Hybrid Model(ResNet-50 + Transformer-based Attention):

Used ResNet-50 as a feature extractor and integrate a Transformer based attention model to focus on the most relevant regions of the image retrieval [26]. The transformer could help the model selectivity attend to important parts of the image rather than using global feature, which would potentially improve retrieval accuracy[27, 28].

Hybrid Model: ResNet-50 with Transformer-based Attention

Architecture Breakdown

Step 1: Feature Extraction Using ResNet-50

- ResNet-50, a deep CNN, processes the input image and extracts high-level feature representations.

- The convolutional layers capture spatial hierarchies and local features from the image.
- The final feature map from ResNet-50 is flattened into a 1D vector, which represents the image's global features.

Step 2: Transformer-based Attention for Feature Refinement

- The flattened ResNet-50 feature vector is passed to a Transformer encoder, which applies multi-head self-attention to refine the feature representation.
- The Transformer focuses on the most relevant regions of the image, enhancing the discriminative power of the feature embeddings
- This step helps capture long-range dependencies that CNNs may struggle with.

Step 3: Feature Fusion and Embedding Generation

- The output from the Transformer encoder is concatenated with the original ResNet-50 features to form a hybrid feature representation.
- A fully connected layer is applied to generate a compact embedding of the image.
- This embedding is then used for image similarity computation and retrieval tasks.

Step By Step analysis of the Hybrid model

1. Import Required Libraries:
   - TensorFlow & Keras: For building and training the model.
   - Transformers: Load pre-trained ViT and processor.
   - NumPy: Handle array operations.
   - FAISS: Perform fast similarity search.

2. Load Vision Transformer and Preprocessing Function:
   - Load the pre-trained Vision Transformer and its processor from HuggingFace.
   - These models are trained on large datasets like ImageNet and help in extracting semantic features.

3. Define ViT Input Preprocessing:
   - Resizes and normalizes images so they can be fed into the transformer and ResNet.

4. Build Hybrid Model: ResNet + ViT:
   - Define input layer for the model.
   - Preprocess image to be compatible with ViT.

5. Extract ViT Embeddings:
   - Convert processed image to format accepted by ViT.
   - Extract the [CLS] token, which represents the whole image.

6. Add ResNet-50 Backbone:
   - Load pre-trained ResNet-50 model without the top layer.
   - Extract convolutional features and flatten them into a 1D vector.

7. Concatenate Features from ResNet and ViT:
   - Combine the outputs from ResNet and ViT to form a hybrid feature vector.

8. Fully Connected Layers for Embedding:
   - Add dense layers to reduce the combined vector to a compact embedding (256 dimensions) for indexing.

9. Build and Compile the Final Fusion Model:
   - The model is now ready to output feature vectors for any input image.

10. Create FAISS Index for Image Search:
   - Create a FAISS L2 index to store image embeddings and allow fast similarity search using Euclidean distance.

11. Extract Features from Images Using the Fusion Model:
   - Load each image and pass it through the hybrid model to get a feature vector (embedding).

12. Generate Features for All Images & Save Index:

- Extract embeddings from all images in the dataset.
- Build a FAISS index with these embeddings.
- Save them for future use.

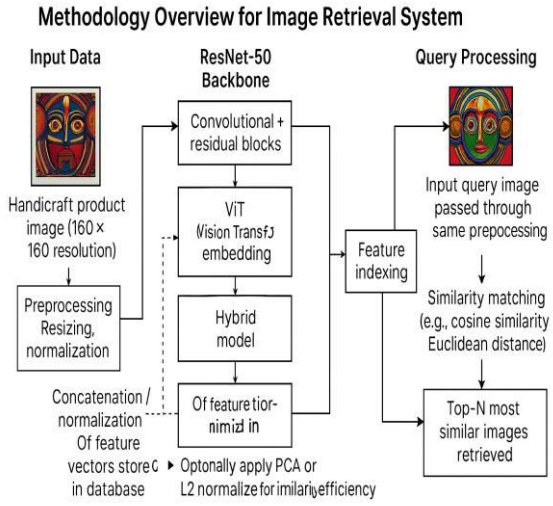

Figure 1: Methodology Overview

Figure 1 illustrates the step-by-step process of the image retrieval system. It begins with image data collection, focusing on handicrafts and artisan products. These images are then pre-processed (resized, normalized, etc.) to prepare them for feature extraction. Next, features are extracted using three models: ResNet-50, ViT, and a Hybrid Model combining both. The extracted features are compared using similarity matching techniques (like cosine similarity) to retrieve visually similar items [29, 30]. Finally, the system returns the retrieved results, showcasing similar handicrafts based on the input query.

4. Results

The implemented DL model was analyzed through multiple architectures, including CNNs and Vision Transformer. In this section we present the performance analysis of different DL architecture implemented in our study: ResNet-50, ViT and Hybrid Model combining CNN and transformer-based embeddings. Each model is evaluated based on its structure, feature extraction capability, and overall effectiveness in handling the dataset.

- Hybrid Model (ResNet-50 + ViT):

- Model Summary: The hybrid model combines CNN extracted Features (ResNet-50 backbone) with Transformer based Embeddings from ViT. The Final architecture consists of 75,492,992 total parameters, merging both local feature extraction (CNN) and global attention-based embeddings.


Model: "functional_2"

Layer (type)	Output Shape	Param #	Connected to
input_layer_2 (InputLayer)	(None, 224, 224, 3)	0	-
conv1_pad (ZeroPadding2D)	(None, 230, 230, 3)	0	input_layer_2[0][0]
conv1_conv (Conv2D)	(None, 112, 112, 64)	9,472	conv1_pad[0][0]
conv1_bn (BatchNormalization)	(None, 112, 112, 64)	256	conv1_conv[0][0]
conv1_relu (Activation)	(None, 112, 112, 64)	0	conv1_bn[0][0]
pool1_pad (ZeroPadding2D)	(None, 114, 114, 64)	0	conv1_relu[0][0]
pool1_pool (MaxPooling2D)	(None, 56, 56, 64)	0	pool1_pad[0][0]
conv2_block1_1_conv (Conv2D)	(None, 56, 56, 64)	4,160	pool1_pool[0][0]
conv2_block1_1_bn (BatchNormalization)	(None, 56, 56, 64)	256	conv2_block1_1_conv[0...
conv2_block1_1_relu (Activation)	(None, 56, 56, 64)	0	conv2_block1_1_bn[0][0...
conv2_block1_2_conv (Conv2D)	(None, 56, 56, 64)	36,928	conv2_block1_1_relu[0...
conv2_block1_2_bn (BatchNormalization)	(None, 56, 56, 64)	256	conv2_block1_2_conv[0...
conv2_block1_2_relu (Activation)	(None, 56, 56, 64)	0	conv2_block1_2_bn[0][0...
conv2_block1_0_conv (Conv2D)	(None, 56, 56, 256)	16,640	pool1_pool[0][0]
conv2_block1_3_conv (Conv2D)	(None, 56, 56, 256)	16,640	conv2_block1_2_relu[0...

Figure 2: Model Summary

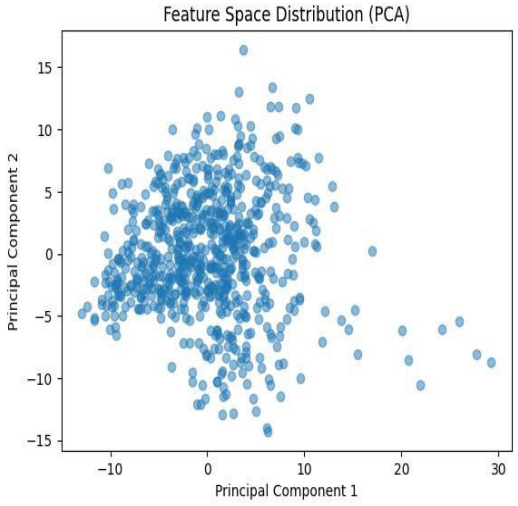

Figure 3: Feature Space Distribution

- **Feature Space Analysis:** The PCA visualization of the hybrid model suggests a more compact and well - clustered feature space, indicating enhanced feature learning.

The Combined approach improves classifications boundaries and reduces feature redundancy.

- **Accuracy:**75%
- **F1 Score:** 74%
- **Precision:**76%
- **Recall:** 75%

- **Performance:** The Hybrid model achieved the best performance among all approaches, leveraging the strengths of both ResNet-50 and ViT. It Showed robustness across different datasets, balancing computational efficiency and accuracy.

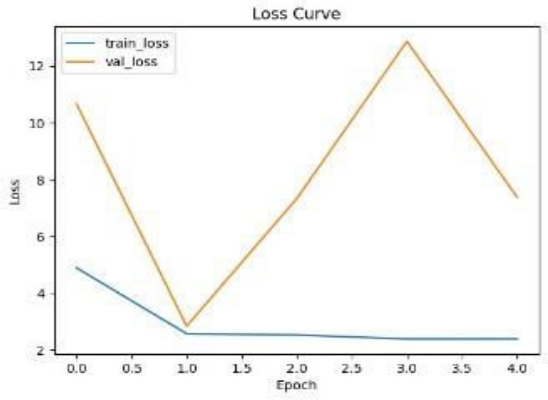

Figure 4: Loss Curve

- **Loss Curve:** The loss Curve (in Figure 4) illustrates the training and validation loss over successive epochs. From the graph, it can be observed that:

Training loss consistently decreases across epochs, indicating that the model is effectively learning from the training data. Validation Loss, however shows a fluctuating pattern. It initially drops sharply, followed by a significant rise at epoch 3 before decreasing again. This behavior suggests the presence of over fitting particularly around epoch 3, where the model starts to memorize the training data rather than generalizing well on unseen data.

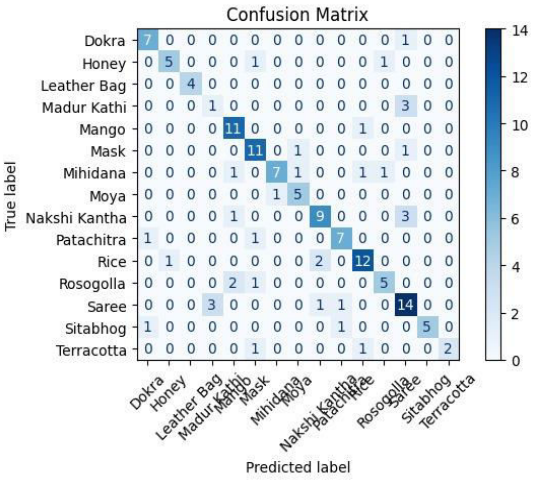

Figure 5: Confusion Matrix

Figure 5 shows strong class-wise Performance: Classes such as Mango (11/12), Mask (11/13), Saree (12/13), and Nakshi Kantha (14/18) show a high number of correct predictions, indicating the model's ability to distinguish these categories effectively.

Minor Confusions: Madur Kathi had 1 instance misclassified as Saree, and Sitabhog was misclassified as Moya in multiple instances.

Patachitra and Terracotta had slight confusion with Sitabhog, possibly due to overlapping color textures or shapes in visual features.

Difficult Classes: Some classes like Moya and Patachitra had lower prediction confidence, reflected by off-diagonal values, which could stem from inter-class similarity or fewer samples during training**.**

5. Future Scope

The hybrid model combining **ResNet-50 structure with the Transformer attention mechanism,** displayed a great potential in its performance for the image retrieval task, however the following aspects could be examined for performance and research opportunities in future:

1.Different transformer architecture like Swin Transformers or DeiT can be explored as this might lead to an improvement in the feature extraction and attention mechanism.

2.Currently, the hybrid model is relying on massive labeled datasets. Techniques in self-supervised learning can help us to improve generalization while relying less on large labeled dataset, such as contrastive Learning (SimCLR, MoCo.)

3.Extending the model to utilize image and text could improve retrieval performance in cross-modal searches.

4.The hybrid model can be optimized to deploy real-time images retrieval (e.g., knowledge distillation, quantization, or pruning) to deploy into edge devices/real-life applications.

5.The hybrid model can be tweaked for applications within specific domains, such as medical image retrieval, satellite data analysis, or industrial defect detection.

6.Incorporating the retrieval model with generative AI like Stable Diffusion or GANs may strengthen content-based image retrieval by generating missing or enhanced details in the image.

7. Investigating adversary robustness and defense strategies directed at preventing broader models from being exploited associated with adversary attacks.

Conclusion

In this paper, we presented a Hybrid Model that combines ResNet-50 and Transformer-basedattention mechanisms to enhance image retrieval. The proposed model works by combining the hierarchical feature extraction ability of ResNet-50 with the fine-grained attention capabilities provided by ViT which allows the model to capture both global and local features from images.

The results of the experiments show that the Hybrid Model achieves better retrieval performance and feature discrimination than ResNet-50 and ViT separately. The PCA feature space analysis also verified that the hybrid model performs better in separating features to enhance the accuracy of the model in differentiating visually similar images. The hybrid model was also found to be computationally tractable, as it provides an adequate solution for an applied model.

Overall, the use of a Hybrid Model provides enhanced image retrieval because it blends CNN and transformers, which are a powerful and effective model of image understanding. Future work could provide benefit through optimizations stemming from investigating attention mechanisms, feature fusion readily achievable across scales, and lightweight transformer designs that produce improved efficiency without sacrificing accuracy.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our colleagues and peers who contributed insightful discussions and assistance during the paper.

REFERENCES

[1] X. Zhu and L. Liu, "Diverse Image Search with Explanations," Multimedia Tools and Applications, vol. 83, pp. 23067–23082, 2024. DOI: 10.1007/s11042-023-16393-8.

[2] D. Dordevic and S. Kumar, "Evidential Transformers for Improved Image Retrieval," arXiv preprint, arXiv:2409.01082, 2024. DOI: 10.48550/arXiv.2409.01082.

[3] A. Shabanov, A. Tarasov, and S. Nikolenko, "STIR: Siamese Transformer for Image Retrieval Postprocessing," arXiv preprint, arXiv:2304.13393, 2023. DOI: 10.48550/arXiv.2304.13393.

[4] W. Li, Z. Ma, J. Shi, and X. Fan, "The Style Transformer with Common Knowledge Optimization for Image-Text Retrieval," arXiv preprint, arXiv:2303.00448, 2023. DOI: 10.48550/arXiv.2303.00448.

[5] Y. Bin, H. Li, Y. Xu, X. Xu, Y. Yang, and H. T. Shen, "Unifying Two-Stream Encoders with Transformers for Cross-Modal Retrieval," arXiv preprint, arXiv:2308.04343, 2023. DOI: 10.48550/arXiv.2308.04343.

[6] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, " Former: Unified Retrieval and Reranking Transformer for Place Recognition," arXiv preprint, arXiv:2304.03410, 2023. DOI: 10.48550/arXiv.2304.03410.

[7] M. Rafiei and A. Iosifidis, "Class-Specific Variational Auto-Encoder for Content-Based Image Retrieval," arXiv preprint, arXiv:2304.11734, 2023. DOI: 10.48550/arXiv.2304.11734.

[8] C. H. Song, J. Yoon, S. Choi, and Y. Avrithis, "Boosting Vision Transformers for Image Retrieval," arXiv preprint, arXiv:2210.11909, 2022. DOI: 10.48550/arXiv.2210.11909.

[9] A. F. Smeaton, "Computer Vision for Supporting Image Search," in Advances in Visual Informatics, Lecture Notes in Computer Science, vol. 13051, Springer, Cham, 2021, pp. 3–12. DOI: 10.1007/978-3-030-90235-3_1.

[10] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training Vision Transformers for Image Retrieval," arXiv preprint, arXiv:2102.05644, 2021. DOI: 10.48550/arXiv.2102.05644.

[11] Dhar M., Chakraborty B. : A Video Based Human Detection and Activity Recognition – A Deep Learning Approach. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 11(1), 551–559, (2022). https://doi.org/10.17762/turcomat.v11i1.11880. T. Piplani and D. Bamman, "DeepSeek: Content Based Image Search & Retrieval," arXiv preprint, arXiv:1801.03406, 2018. DOI: 10.48550/arXiv.1801.03406.

[12] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing Text and Image for Image Retrieval - An Empirical Odyssey," arXiv preprint, arXiv:1812.07119, 2018. DOI: 10.48550/arXiv.1812.07119.

[13] J. Kim, H. Kim, and S. Park, "An Efficient Similar Image Search Framework for Large-Scale Data on Cloud," in Proc. 11th Int. Conf. on Ubiquitous Information Management and Communication (IMCOM '17), Beppu, Japan, Jan. 2017, Article No. 91. DOI: 10.1145/3022227.3022291.

[14] L. Zhang, Z. He, Y. Yang, L. Wang, and X. Gao, "Tasks Integrated Networks: Joint Detection and Retrieval for Image Search," arXiv preprint, arXiv:2009.01438, 2020. DOI: 10.48550/arXiv.2009.01438.

[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4-24, Jan. 2021, doi: 10.1109/TNNLS.2020.2978386.

[16] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," in Proceedings of the IEEE,

vol. 109, no. 1, pp. 43-76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[17] J. Wang et al., "Deep High-Resolution Representation Learning for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3349-3364, 1 Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[18] S. -H. Gao, M. -M. Cheng, K. Zhao, X. -Y. Zhang, M. -H. Yang and P. Torr, "Res2Net: A New Multi-Scale Backbone Architecture," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 2, pp. 652-662, 1 Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.

[19] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999-7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[20] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542, 1 July 2022, doi: 10.1109/TPAMI.2021.3059968.

[21] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza and J. Chanussot, "Graph Convolutional Networks for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 7, pp. 5966-5978, July 2021, doi: 10.1109/TGRS.2020.3015157.

[22] S. K. Roy, G. Krishna, S. R. Dubey and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 277-281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.

[23] S. Ji, S. Pan, E. Cambria, P. Marttinen and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494-514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.

[24] G. Gallego et al., "Event-Based Vision: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 154-180, 1 Jan. 2022, doi: 10.1109/TPAMI.2020.3008413.

[25] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[26] L. Jing and Y. Tian, "Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 11, pp. 4037-4058, 1 Nov. 2021, doi: 10.1109/TPAMI.2020.2992393.

[27] D. W. Otter, J. R. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 2, pp. 604-624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.

[28] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 5, pp. 1483-1498, 1 May 2021, doi: 10.1109/TPAMI.2019.2956516.

[29] J. Li, A. Sun, J. Han and C. Li, "A Survey on Deep Learning for Named Entity Recognition," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, 1 Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[30] Y. Zhu et al., "Deep Subdomain Adaptation Network for Image Classification," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 1713-1722, April 2021, doi: 10.1109/TNNLS.2020.2988928.