SIM-Trans: Structure Information Modeling Transformer for Fine-grained Visual Categorization

Hongbo Sun, Xiangteng He, and Yuxin Peng*

Wangxuan Institute of Computer Technology, Peking University

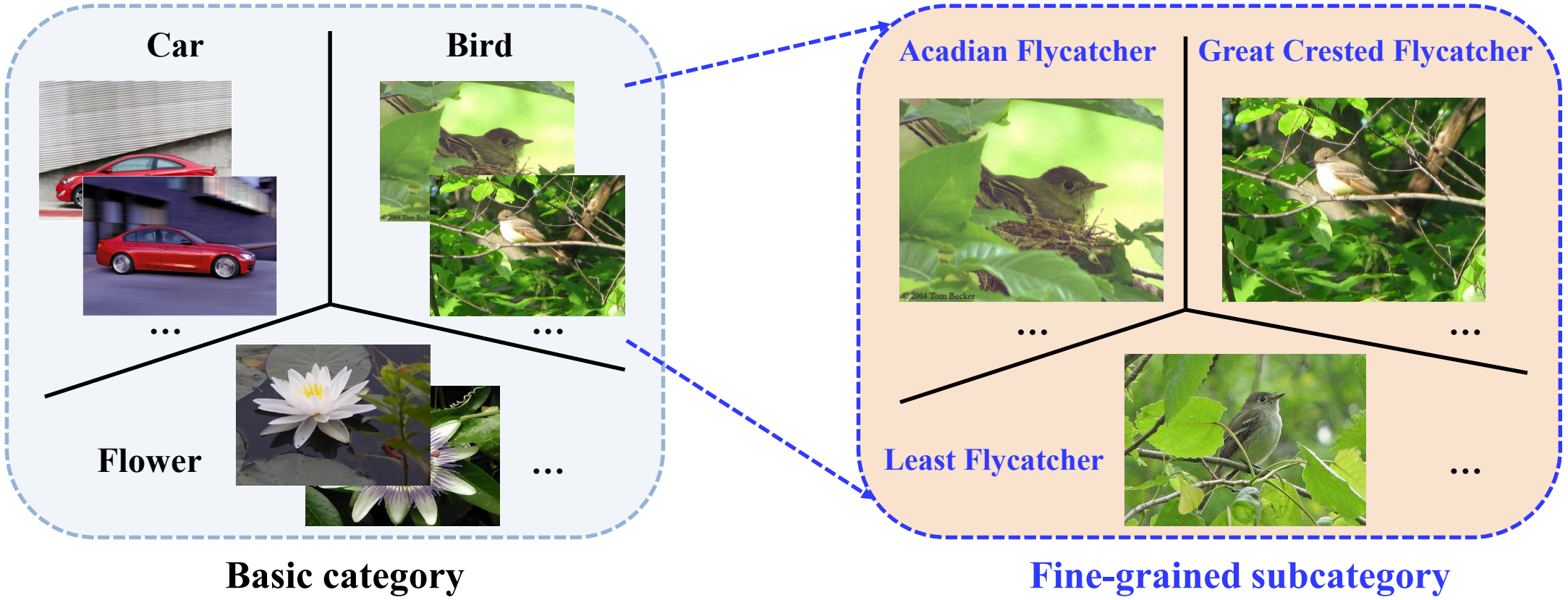
{sunhongbo, hexiangteng, pengyuxin}@pku.edu.cn

Outline

- Introduction
- Our Approach
- Experiment
- Conclusion

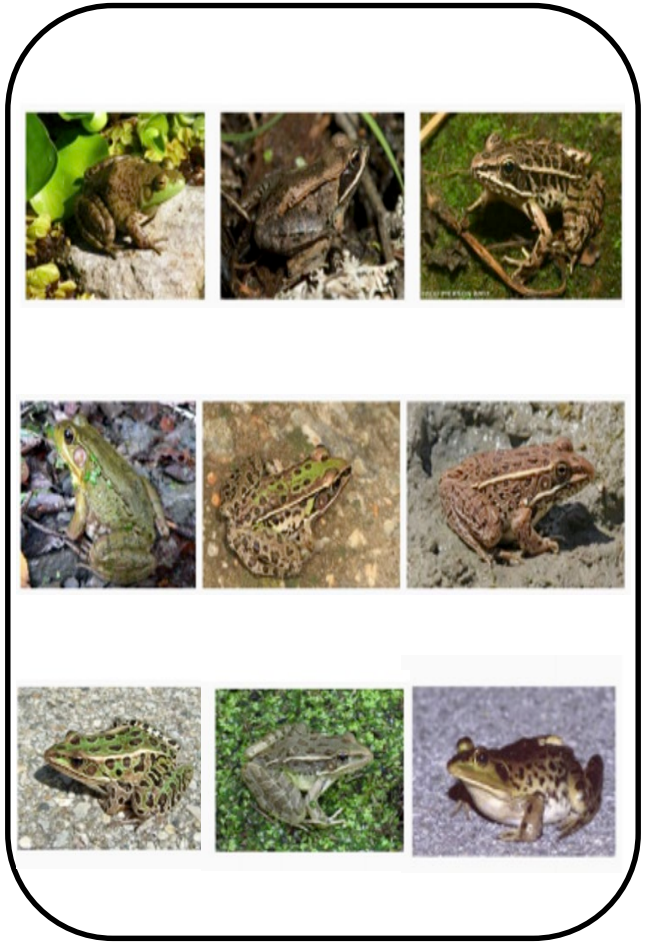
Background

- **Fine-grained Visual Categorization**
  - **Recognize** object into **fine-grained subcategories** from a given **basic category**, such as identifying bird species

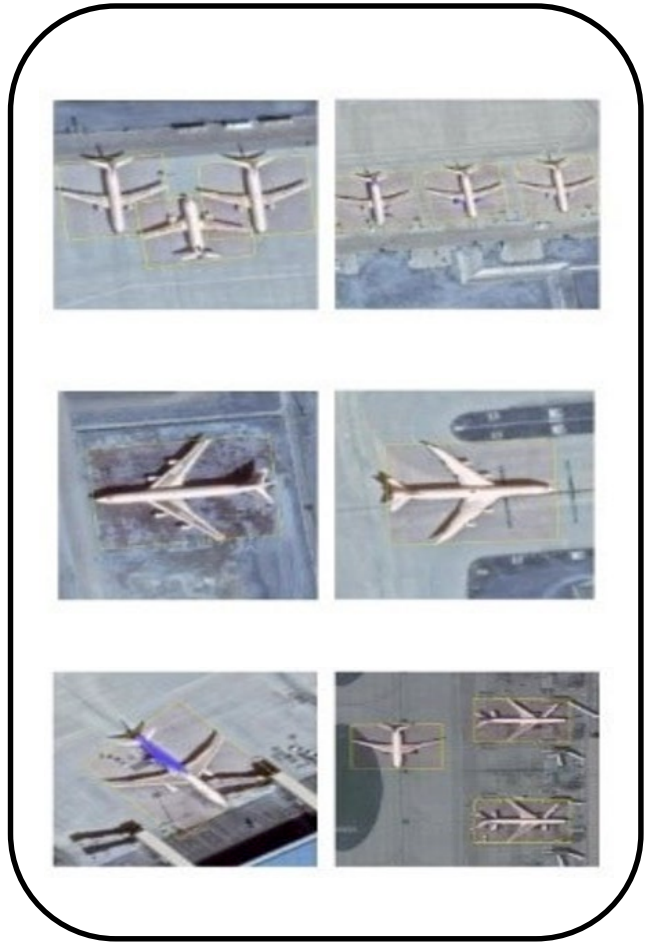


Application Scenarios

- **Biodiversity conservation**



- **Remote sensing object recognition**



- **Smart Retail**

Research Challenages

- **Large Intra-class Variance**
- **Small Inter-class Variance**

Heermann Gull



Slaty backed Gull



Western Gull



Research Challenages

- Large Intra-class Variance
- Small Inter-class Variance

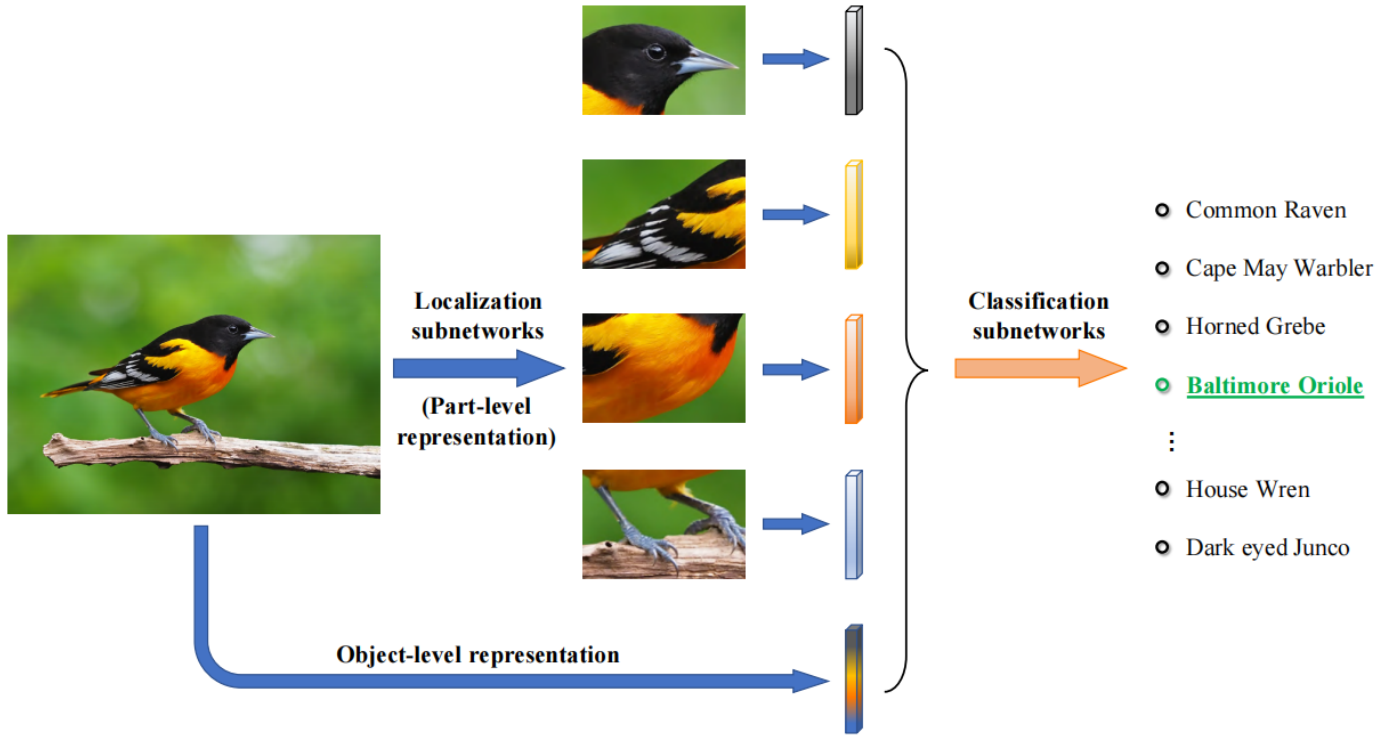
Heermann Gull

Slaty backed Gull

Western Gull

Related Work

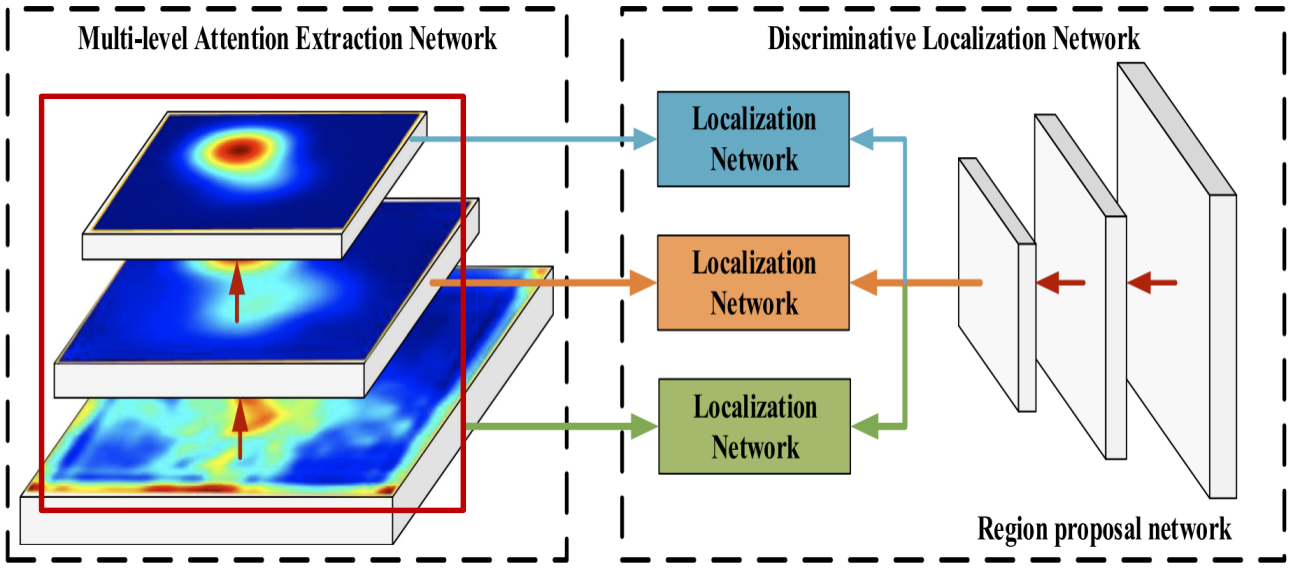
- **Mainly adopt localization and recognition paradigm (CNN based method)**
  - Utilize detection or segmentation method
  - Leverage attention mechanism
  - Use deep filters

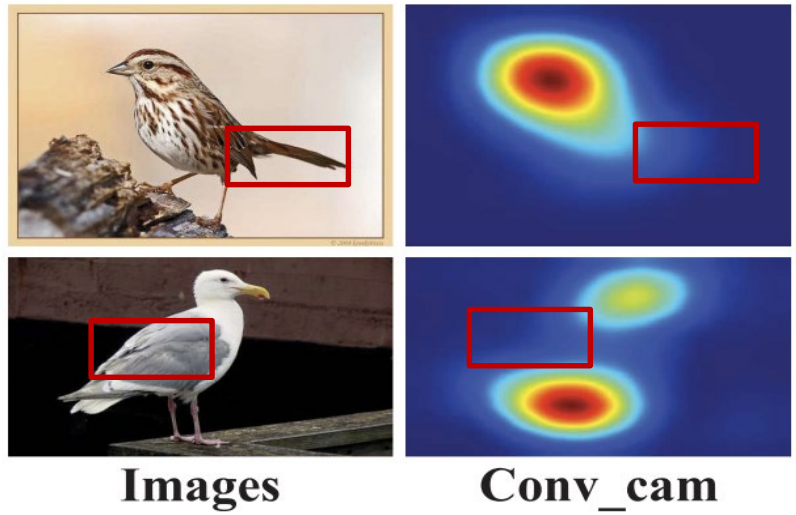


Wei et al., Fine-Grained Image Analysis with Deep Learning: A Survey, TPAMI 2021.

Related Work

- **Mainly adopt localization and recognition paradigm (CNN based method)**
  - Spatial details information degrades with stacked convolution and pooling operation
  - Significant regions are generally hard to detect fully within the object extent

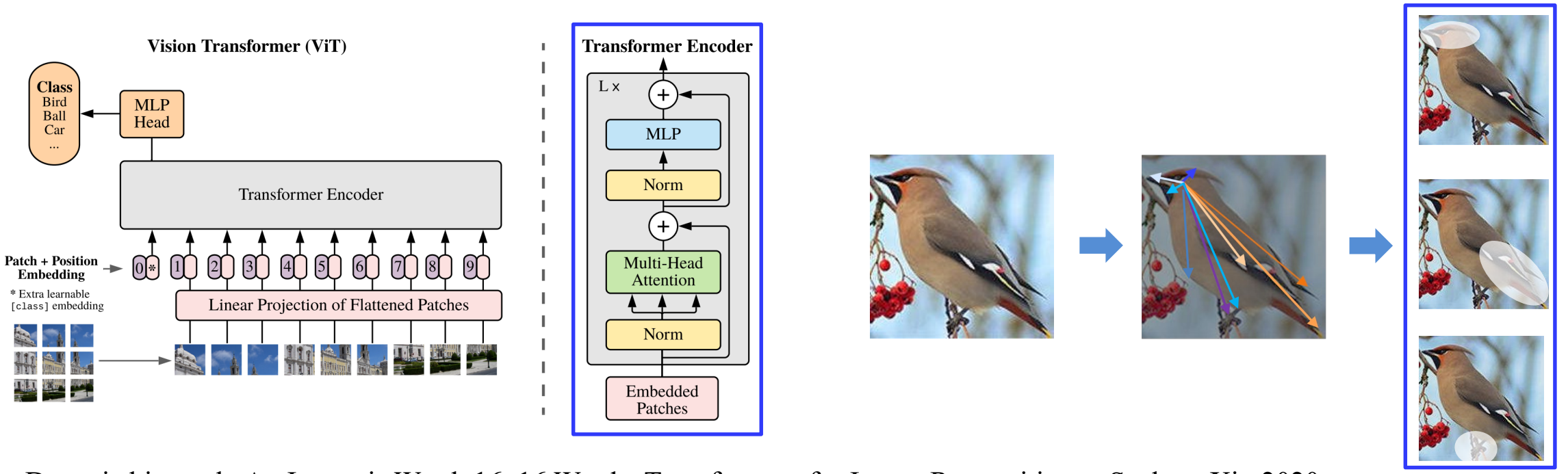




He et al., Fast Fine-Grained Image Classification via Weakly Supervised Discriminative Localization, TCSVT 2019.

Motivation

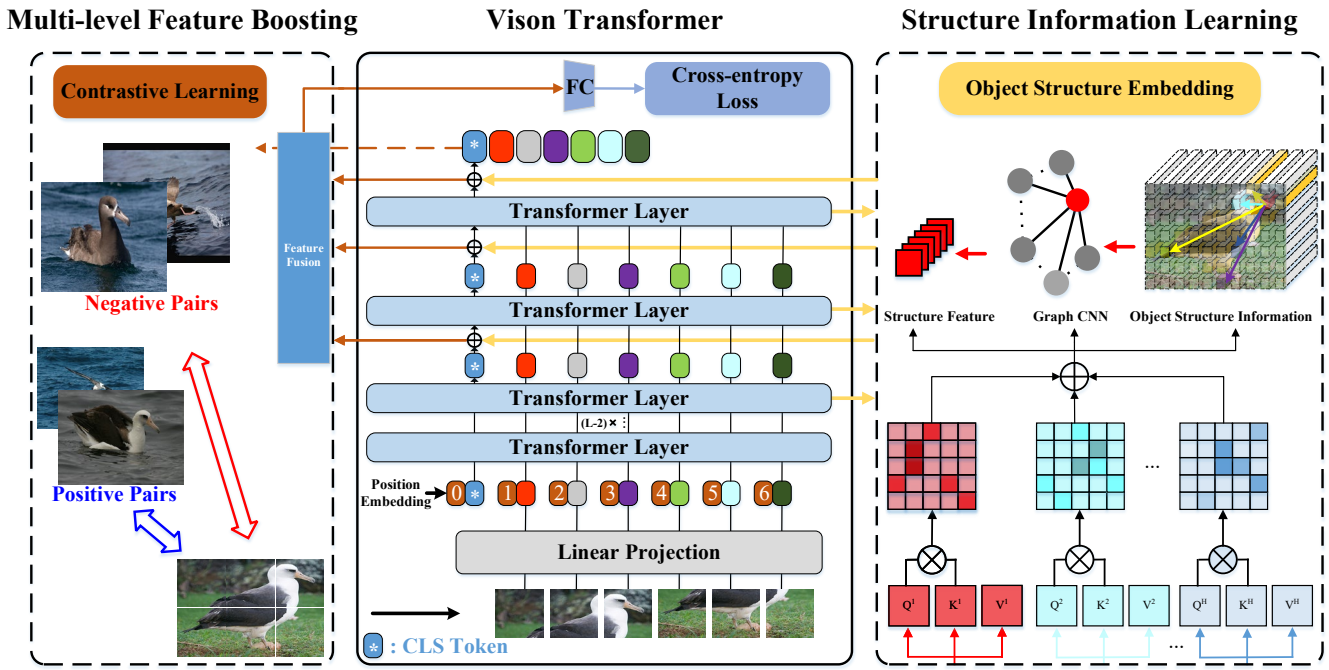
- **Vision transformer provides interaction among patches layer by layer**
  - Spatial details information is kept via patch information interaction in the transformer layer
- **Structure reveals the object's spatial composition of significant regions**
  - Helpful for highlighting significant regions within object for fine-grained recognition



Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv 2020.

Our Contribution

- **We propose a Structure Information Modeling Transformer (SIM-Trans) approach**
  - **Structure information learning** is introduced into vision transformer to mine the object's spatial context relation for **highlighting discriminative regions** within object extent
  - **Multi-level feature boosting** is proposed to utilize the complementarity of multi-layer features and contrastive learning for **obtaining robust feature representation**

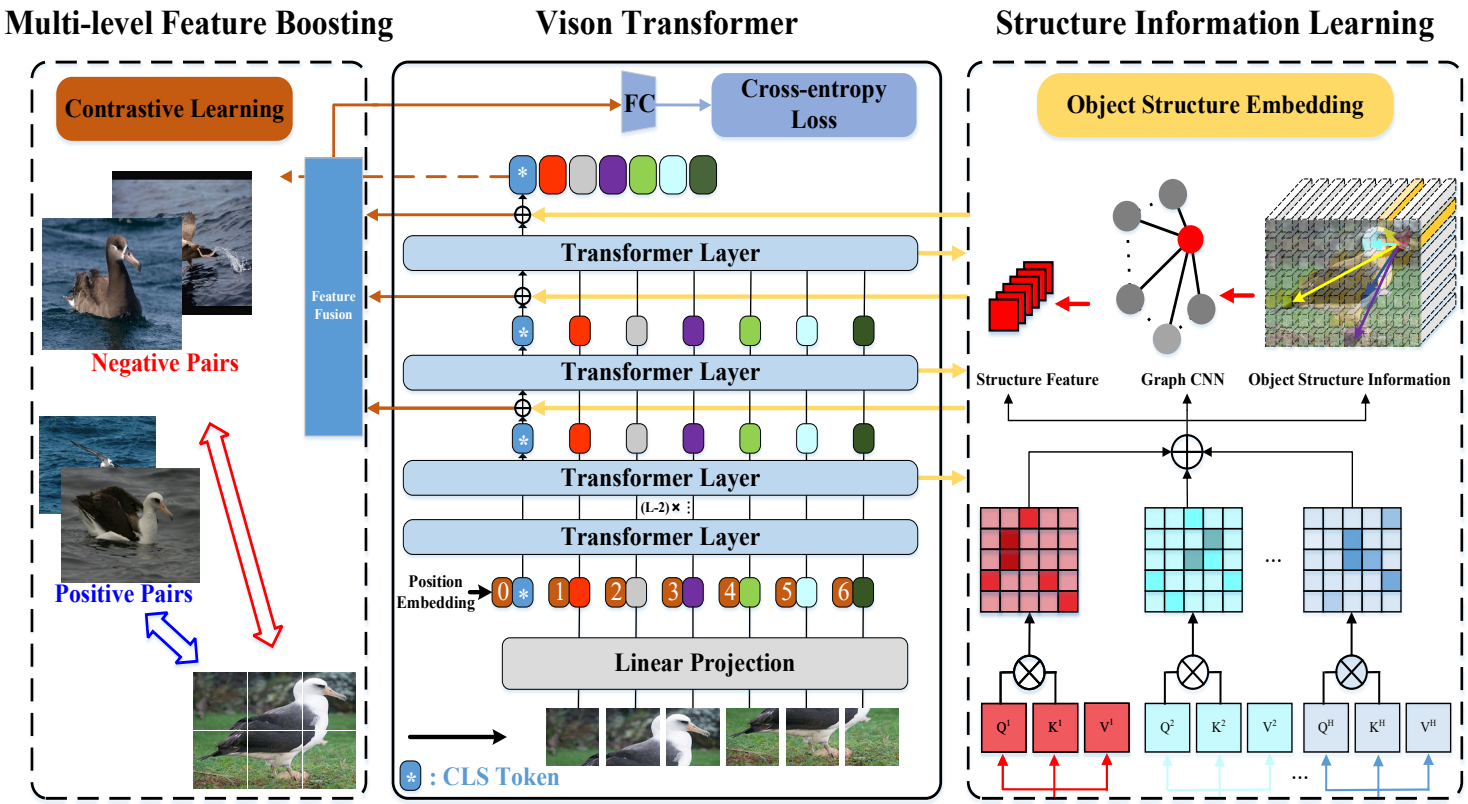


Outline

- Introduction
- **Our Approach**
- Experiment
- Conclusion

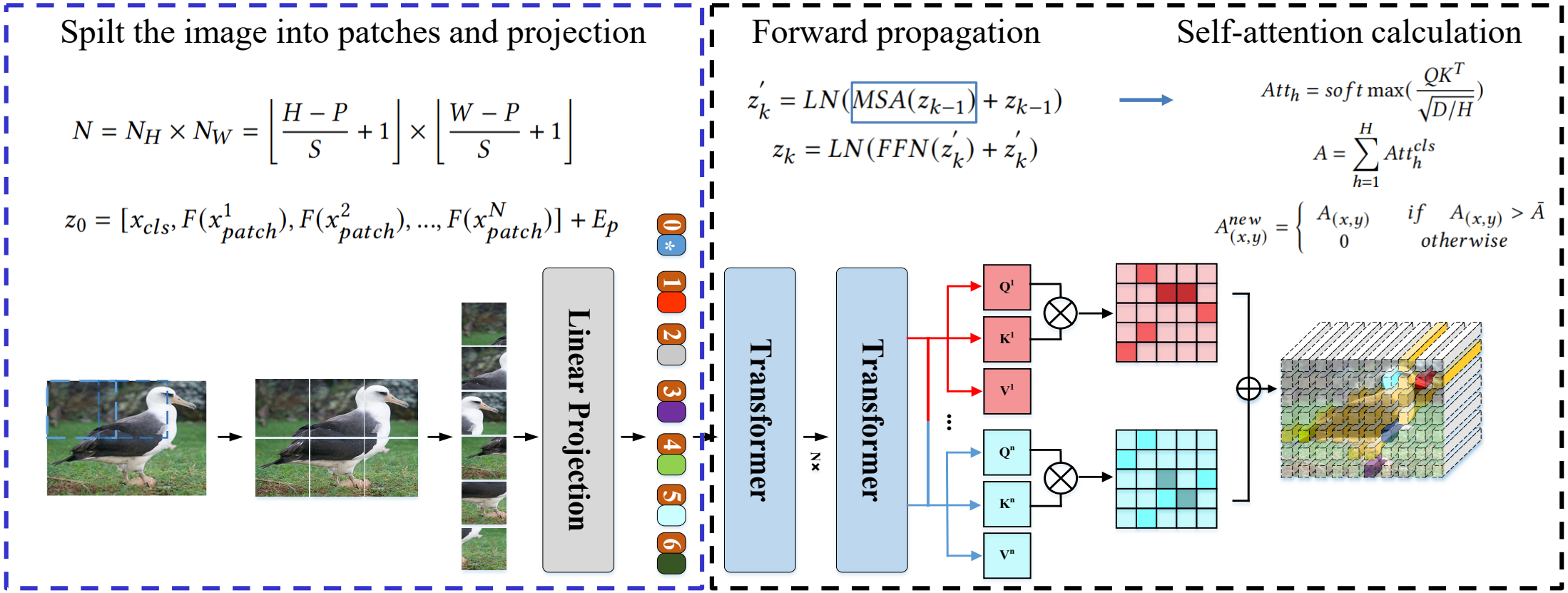
Our Approach

- **Structure Information Learning**
  - Patch significance obtaining
  - Object structure embedding
- **Multi-level Feature Boosting**
  - Contrastive learning
  - Multi-layer features concatenation
- **Final Classification**



Structure Information Learning

- **Patch Significance Obtaining**
  - Adopt sliding window splitting method with overlap to get image patches
  - Obtain patch significance based on self-attention calculation



Structure Information Learning

- **Object structure Embedding**
  - Calculate the spatial context relation and construct structure graph

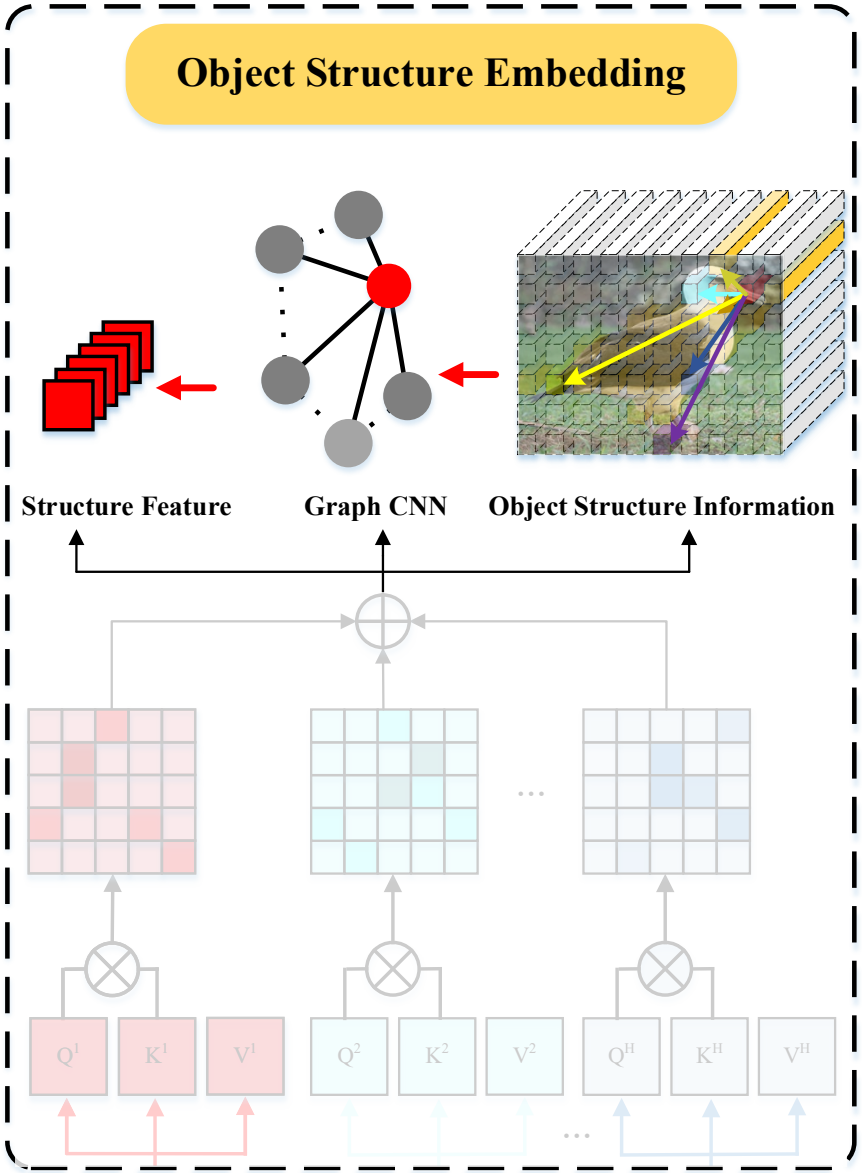
$$\rho_{x,y} = \sqrt{(\frac{x - x_0}{N_W})^2 + (\frac{y - y_0}{N_H})^2}$$

$$\theta_{x,y} = \frac{(\arctan 2(y - y_0, x - x_0) + \pi)}{2\pi}$$

  - Extract structure features by graph convolutional network

$$Adj = A^{new} \times (A^{new})^T$$

$$S = \sigma(Adj \times \sigma(Adj \times X \times W^1) \times W^2)$$



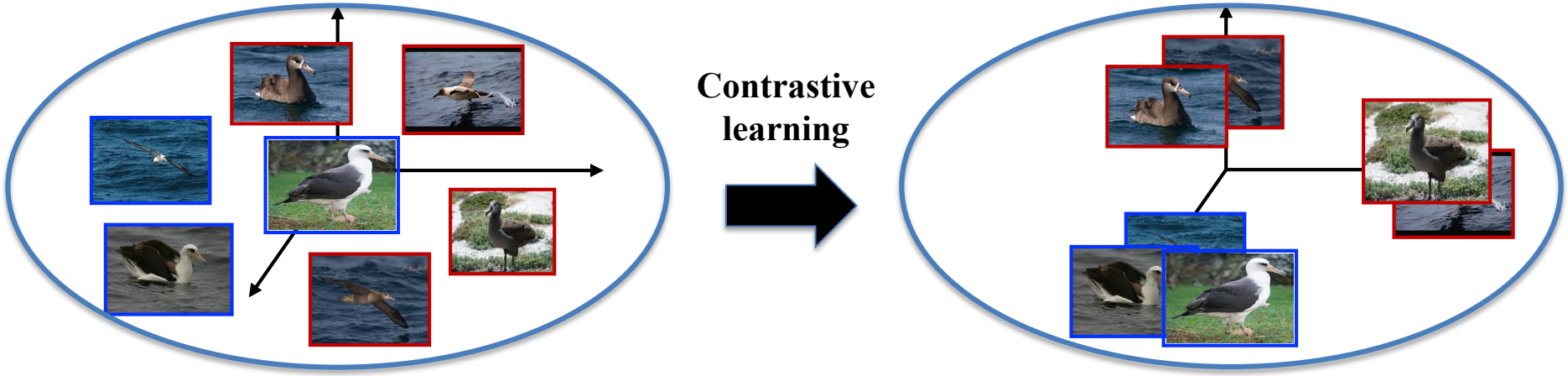
Multi-level Feature Boosting

- **Contrastive Learning**
  - Pull the positive pairs and push the negative pairs in the feature space

Calculate contrastive learning loss

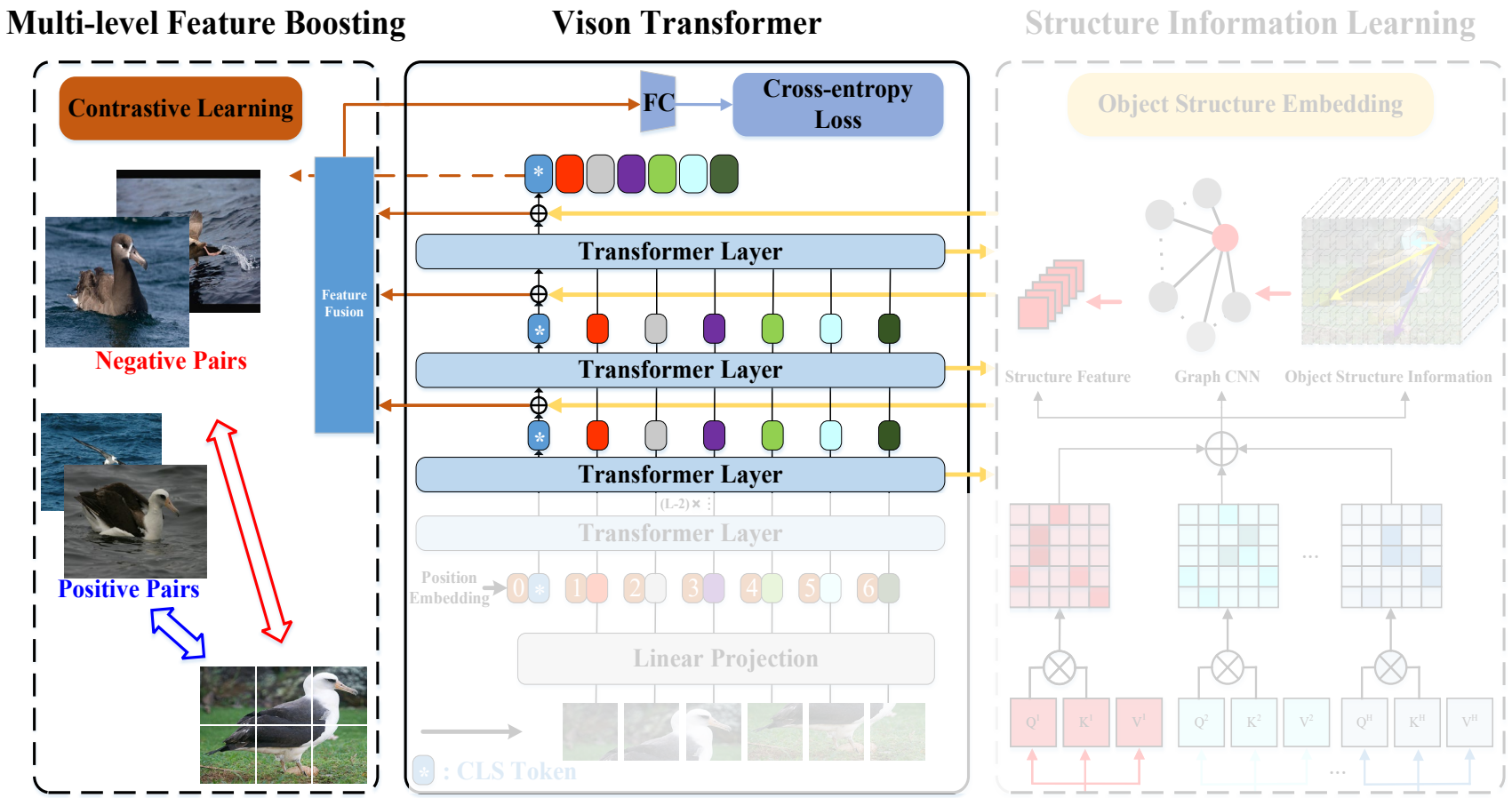
$$Indicator_{i,j} = \max\{0, \alpha + sim(z_i, z_j^-) - \frac{1}{\Gamma_{y(i)=y(j), i\neq j}} \sum_{j:i\neq j} sim(z_i, z_j^+)\}$$

$$L_{CL} = \frac{1}{N^2} \sum_{i=1}^{N} [\sum_{j:y(i)=y(j)}^{N} (1 - sim(z_i, z_j^+)) + \sum_{j:y(i)\neq y(j)}^{N} Indicator_{i,j} \times sim(z_i, z_j^-)]$$



Multi-level Feature Boosting

- **Multi-layer feature concatenation**
  - Exploit the complementarity of multi-layer features to obtain robust representation for recognition



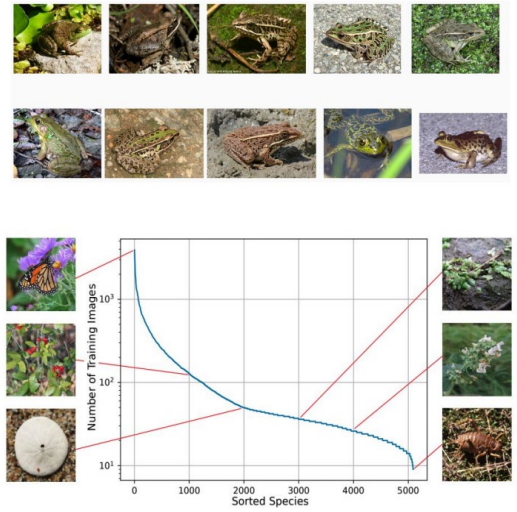
Outline

- Introduction
- Our Approach
- Experiment
- Conclusion

Dataset

- **CUB-200-2011**
  - 200 classes
  - 5994 training images
  - 5794 test images

- **iNaturalist 2017**
  - 13 super classes, 5089 classes
  - 579,184 training images
  - 95,986 test images



Comparisons with State-of-the-art Methods

- **On CUB-200-2011 dataset**
  - Our SIM-Trans approach can achieve better fine-grained recognition performance than CNN backbone based methods

Method	Backbone	Acc(%)
KP (CVPR 2017) [6]	VGG16	86.2
MA-CNN (ICCV 2017) [39]	VGG19	86.5
PC (ECCV 2018) [10]	DenseNet161	86.9
NTS-Net (ECCV 2018) [36]	ResNet50	87.5
M2DRL (IJCV 2019) [15]	VGG16	87.2
S3N (ICCV 2019) [7]	ResNet50	88.5
FDL (AAAI 2020) [21]	ResNet50	88.6
LIO (CVPR 2020) [42]	ResNet50	88.0
PMG (ECCV 2020) [9]	ResNet50	89.6
DP-Net (AAAI 2021) [31]	ResNet50	89.3
GaRD (CVPR 2021) [38]	ResNet50	89.6
Chang et al. (CVPR 2021) [4]	ResNet50	89.9
SPS (ICCV 2021) [17]	ResNet50	88.7
Joung et al. (ICCV 2021) [19]	ResNet50	88.4
CAL (ICCV 2021) [24]	ResNet101	90.6
MCEN (ACM MM 2021) [20]	ResNet50	89.3
ViT (ICLR 2020) [8]	ViT-B_16	90.6
RAMS-Trans (ACM MM 2021) [16]	ViT-B_16	91.3
Our SIM-Trans approach	**ViT-B_16**	**91.8**

Comparisons with State-of-the-art Methods

- **On CUB-200-2011 dataset**
  - Our SIM-Trans approach also achieves better performance than transformer backbone based methods

Method	Backbone	Acc(%)
KP (CVPR 2017) [6]	VGG16	86.2
MA-CNN (ICCV 2017) [39]	VGG19	86.5
PC (ECCV 2018) [10]	DenseNet161	86.9
NTS-Net (ECCV 2018) [36]	ResNet50	87.5
M2DRL (IJCV 2019) [15]	VGG16	87.2
S3N (ICCV 2019) [7]	ResNet50	88.5
FDL (AAAI 2020) [21]	ResNet50	88.6
LIO (CVPR 2020) [42]	ResNet50	88.0
PMG (ECCV 2020) [9]	ResNet50	89.6
DP-Net (AAAI 2021) [31]	ResNet50	89.3
GaRD (CVPR 2021) [38]	ResNet50	89.6
Chang et al. (CVPR 2021) [4]	ResNet50	89.9
SPS (ICCV 2021) [17]	ResNet50	88.7
Joung et al. (ICCV 2021) [19]	ResNet50	88.4
CAL (ICCV 2021) [24]	ResNet101	90.6
MCEN (ACM MM 2021) [20]	ResNet50	89.3
ViT (ICLR 2020) [8]	ViT-B_16	90.6
RAMS-Trans (ACM MM 2021) [16]	ViT-B_16	91.3
Our SIM-Trans approach	**ViT-B_16**	**91.8**

Comparisons with State-of-the-art Methods

- **On iNaturalist 2017 dataset**
  - Our SIM-Trans approach achieves better performance than other state-of-the-art methods
  - Achieve promising performance on large-scale fine-grained recognition scenario

Super Class	Class	Train Images	Test Images
Plantae	2101	158407	38206
Insecta	1021	100479	18076
Aves	964	214295	21226
Reptilia	289	35201	5680
Mammalia	186	29333	3490
Fungi	121	5826	1780
Amphibia	115	15318	2385
Mollusca	93	7536	1841
Animalia	77	5228	1362
Arachnida	56	4873	1086
Actinopterygii	53	1982	637
Chromista	9	398	144
Protozoa	4	308	73
Total	5089	579184	95986

Method	Backbone	Acc(%)
ResNet152 (CVPR 2016) [12]	ResNet152	59.0
IncResNetV2 (AAAI 2017) [28]	InResNetV2	67.3
SSN (ECCV 2018) [25]	ResNet101	65.2
TASN (CVPR 2019) [40]	ResNet101	68.2
Huang et al. (CVPR 2020) [18]	ResNet101	66.8
ViT (ICLR 2020) [8]	ViT-B_16	67.0
RAMS-Trans (ACM MM 2021) [16]	ViT-B_16	68.5
Our SIM-Trans approach	**ViT-B_16**	**69.9**

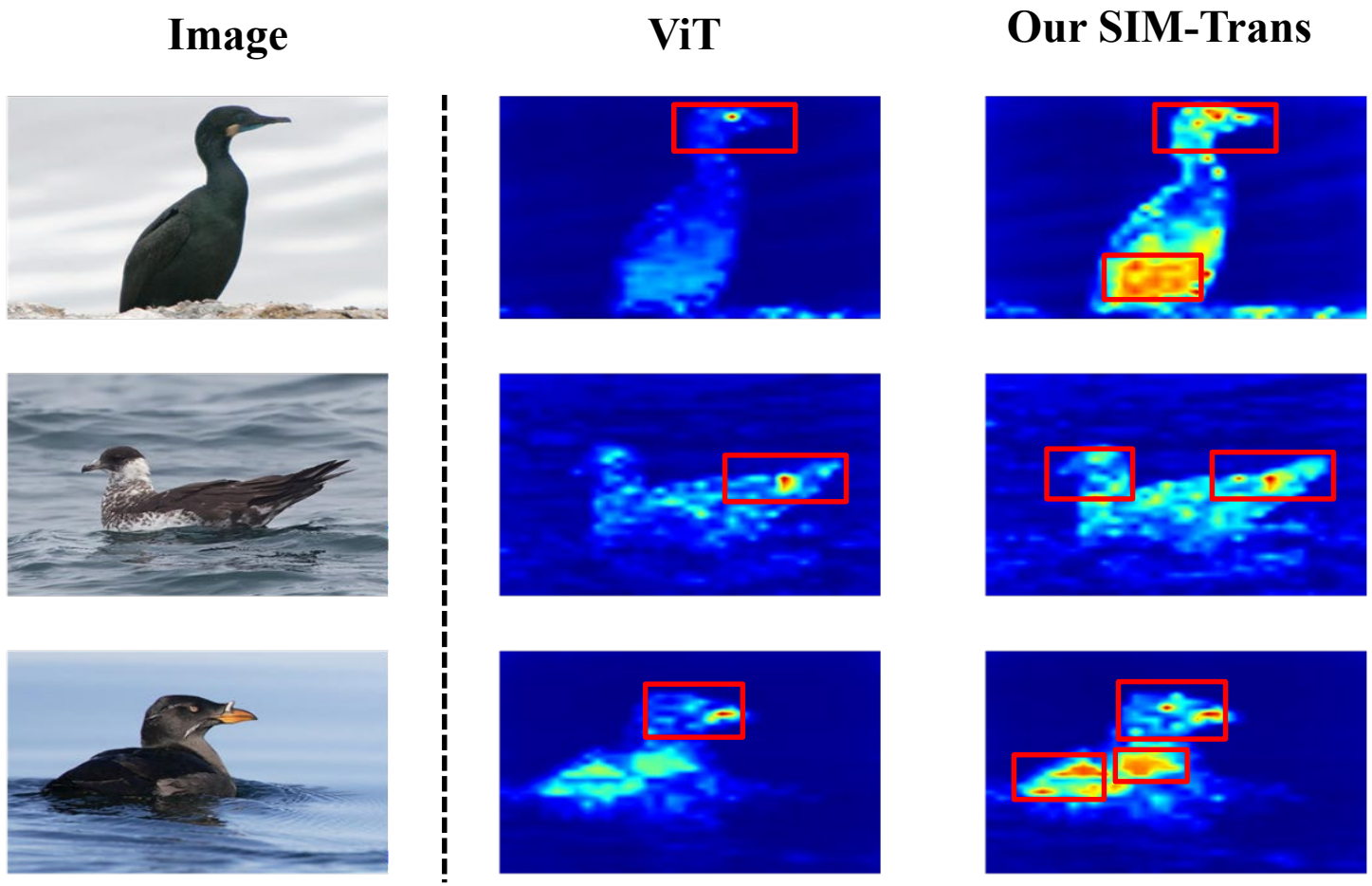
Ablation Study

- **On CUB-200-2011 dataset**
  - Our SIM-Trans approach with the proposed structure information learning (SIL) and multi-level feature boosting (MFB) achieves the best performance
  - The contrastive learning in multi-level feature boosting (MFB) boosts the feature representation robustness to bring performance gains

Method	Acc(%)
Baseline	90.6
Baseline + SIL	91.1
Baseline + SIL + MFB_without_CL	91.4
Baseline + SIL + MFB	**91.8**

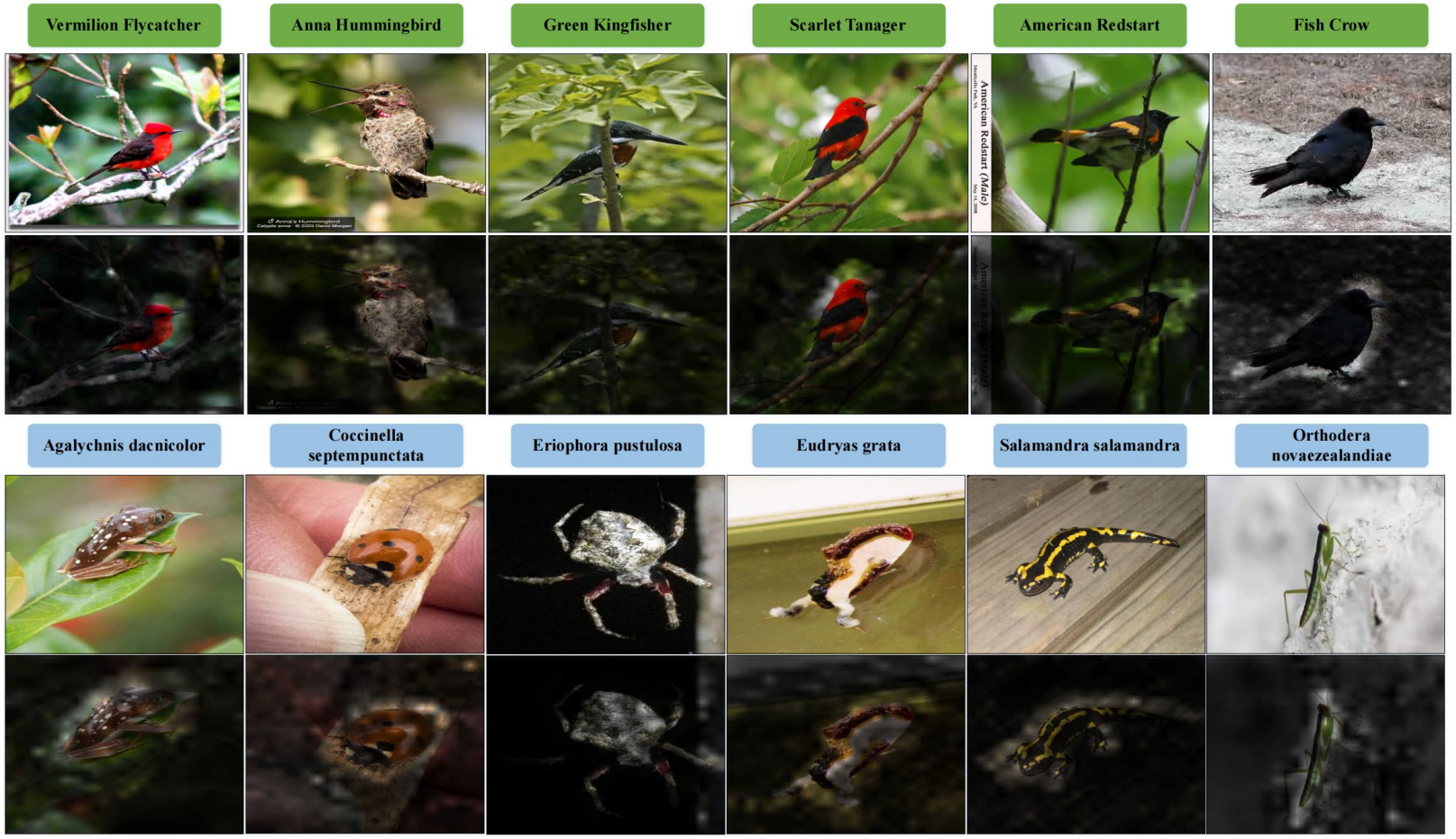
Qualitative Experimental Results

- **Attention Visualization**
  - Our SIM-Trans approach can highlight significant regions within the object more precisely



Qualitative Experimental Results

- **Attention Visualization**
  - Our SIM-Trans approach can highlight significant regions within the object more precisely



Outline

- Introduction
- Our Approach
- Experiment
- Conclusion

Conclusion

- We propose the SIM-Trans approach to introduce the object structure information into vision transformer for boosting the discriminative feature learning to contain both the appearance and structure information

- Structure information learning is proposed to mine spatial context relation of significant patches within the object extent to boost model's understanding ability for object structure and highlight discriminative regions

- Multi-level feature boosting is proposed to exploit the complementarity of multi-layer features and contrastive learning to enhance feature representation robustness

References

- https://www.kaggle.com/c/inaturalist-challenge-at-fgvc-2017
- Sun X, Wang P, Yan Z, et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 184: 116-130.
- Wei X S, Cui Q, Yang L, et al. RPC: A large-scale retail product checkout dataset[J]. arXiv preprint arXiv:1901.07249, 2019.
- Wei X S, Song Y Z, Mac Aodha O, et al. Fine-grained image analysis with deep learning: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
- He X, Peng Y, Zhao J. Fast fine-grained image classification via weakly supervised discriminative localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(5): 1394-1407.
- Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. 2011.
- Van Horn G, Mac Aodha O, Song Y, et al. The inaturalist species classification and detection dataset[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8769-8778.
- Chang D, Pang K, Zheng Y, et al. Your" Flamingo" is My" Bird": Fine-Grained, or Not[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11476-11485.
- Cui Y, Zhou F, Wang J, et al. Kernel pooling for convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2921-2930.
- Ding Y, Zhou Y, Zhu Y, et al. Selective sparse sampling for fine-grained image recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6599-6608.

References

- Du R, Chang D, Bhunia A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches[C]//European Conference on Computer Vision. Springer, Cham, 2020: 153-168.
- Dubey A, Gupta O, Guo P, et al. Pairwise confusion for fine-grained visual classification[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 70-86.
- He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- He X, Peng Y, Zhao J. Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization[J]. International Journal of Computer Vision, 2019, 127(9): 1235-1255.
- Hu Y, Jin X, Zhang Y, et al. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4239-4248.
- Huang S, Wang X, Tao D. Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 620-629.
- Huang Z, Li Y. Interpretable and accurate fine-grained recognition via region grouping[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8662-8672.
- Joung S, Kim S, Kim M, et al. Learning canonical 3d object representation for fine-grained recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1035-1045.
- Li G, Wang Y, Zhu F. Multi-branch Channel-wise Enhancement Network for Fine-grained Visual Recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 5273-5280.
- Liu C, Xie H, Zha Z J, et al. Filtration and distillation: Enhancing region attention for fine-grained visual categorization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11555-11562.

References

- Rao Y, Chen G, Lu J, et al. Counterfactual attention learning for fine-grained visual categorization and re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1025-1034.
- Recasens A, Kellnhofer P, Stent S, et al. Learning to zoom: a saliency-based sampling layer for neural networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 51-66.
- Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI conference on artificial intelligence. 2017.
- Zhou M, Bai Y, Zhang W, et al. Look-into-object: Self-supervised structure modeling for object recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11774-11783.
- Wang S, Li H, Wang Z, et al. Dynamic Position-aware Network for Fine-grained Image Recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(4): 2791-2799.
- Yang Z, Luo T, Wang D, et al. Learning to navigate for fine-grained classification[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 420-435.
- Zhao Y, Yan K, Huang F, et al. Graph-based high-order relation discovery for fine-grained recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15079-15088.
- Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5209-5217.
- Zheng H, Fu J, Zha Z J, et al. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5012-5021.

Contact

Lab Homepage

Github Homepage

***M**ultimedia **I**nformation **P**rocessing **L**ab* (*MIPL*)

http://www.wict.pku.edu.cn/mipl/

SIM-Trans code: https://github.com/PKU-ICST-MIPL/SIM-Trans_ACMMM2022