Graph Neural Networks in Computer Vision - Architectures, Datasets and Common Approaches

Maciej Krzywda*, Szymon Łukasik*‡, Amir H. Gandomi†

*Faculty of Physics and Applied Computer Science, AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: krzywda@agh.edu.pl, slukasik@agh.edu.pl
†Faculty of Engineering and IT, University of Technology Sydney
5 Broadway, Ultimo NSW 2007, Australia
Email: gandomi@uts.edu.au
‡Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: slukasik@ibspan.waw.pl

Abstract—**Graph Neural Networks (GNNs) are a family of graph networks inspired by mechanisms existing between nodes on a graph. In recent years there has been an increased interest in GNN and their derivatives, i.e., Graph Attention Networks (GAT), Graph Convolutional Networks (GCN), and Graph Recurrent Networks (GRN). An increase in their usability in computer vision is also observed. The number of GNN applications in this field continues to expand; it includes video analysis and understanding, action and behavior recognition, computational photography, image and video synthesis from zero or few shots, and many more. This contribution aims to collect papers published about GNN-based approaches towards computer vision. They are described and summarized from three perspectives. Firstly, we investigate the architectures of Graph Neural Networks and their derivatives used in this area to provide accurate and explainable recommendations for the ensuing investigations. As for the other aspect, we also present datasets used in these works. Finally, using graph analysis, we also examine relations between GNN-based studies in computer vision and potential sources of inspiration identified outside of this field.**

Index Terms—**Graph Neural Networks (GNNs), Graph Convolutional Networks (GCNs), Computer Vision**

I. Introduction

The area of computer vision is a very vast area, with its application spanning from the face identification of mobile phone users to the analysis of satellite images. It also employs a plethora of different techniques, based either on statistical apparatus or unconventional nature-inspired algorithms, such as neural networks. This paper aims to consolidate and analyze current knowledge on the application of Graph-Neural Networks in computer vision. We aim to identify the most common architectures and benchmark datasets. By studying relations between different studies, we also try to identify sources of inspiration stemming from other machine learning domains.

The paper is organized as follows, first, in Section , a brief overview of Graph Neural Networks and the most common architectures is provided. Section III covers significant contributions in this field, providing a review of modifications used in the schemes of GNNs. In the next Section IV the most popular datasets used as a point of reference for experimental studies are listed. The relations between identified contributions to the field of computer vision are studied in Section V. It also captures major concepts derived from other sources, including papers outside of the computer vision domain. The paper concludes with summarizing comments and insights gathered during our studies.

II. Graph Neural Networks and Their Basic Forms

Graph Neural Networks (GNNs) [1] can model the relationship of nodes of the graph and generate its numerical representation. It can help to express and provide a more explainable representation of such data. Graphs can represent product or customer similarities in recommendation systems, agent interactions in multiagent robotics, or transceivers in a wireless communication network. Computer vision graphs, primarily formulated as Graph Neural Networks, can be an effective tool for biometrics, face, gesture, body pose recognition on video or images. In this paper, we are using GNN as the root of graphs networks family in which other features like convolutional layers or attention mechanisms can also be found.

A. Graph Convolutional Networks

Graph Convolutional Network (GCNs) [2] [3] [4] is a kind of convolutional neural network that can work directly with graphs and their structural information. Instead of having a 2D/3D array as input in convolutional neural networks, GCNs takes a graph as an input. Similar to how convolutional neural networks extract the essential information from an image to classify the image, GCNs also pass a filter over a graph, searching for important vertices and edges that can be used to classify nodes within the graph.

Figure 1 demonstrates the simplified architecture of Graph Convolutional Networks that is made up of the hidden layers, and it is performing the three steps preparation, aggregation, and update. It works in sequence: input node representations
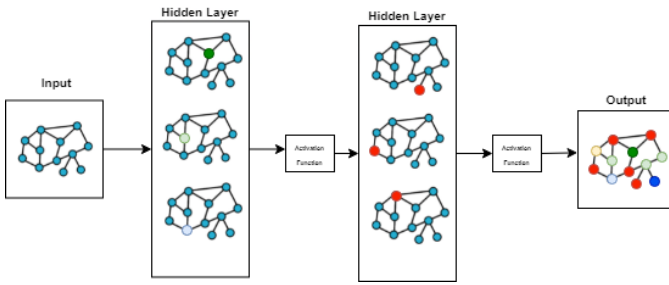
Fig. 1. Graph Convolutional Network with two hidden layers.

are processed using a feedforward network module to produce a message, next the messages of the neighbors of each node are aggregated with respect to the edge weights, and finally, updates are combined and processed to produce the new state of the node representations.

B. Graph Recurrent Networks

Graph Recurrent Networks (GRNs), based on the simple Recurrent Neural Networks with two dense networks, achieve similar performances with the existing models for classifying the nodes on graphs. GRNs uses the random-walk transition matrix to generate the matrices sequence, feeds the output tensor into the classification and the dense attention networks, respectively, and computes the score for the labels. GRNs take a hidden state for each graph node, and it relies on an iterative message passing framework to update these hidden states in parallel. Within each iteration, neighboring nodes exchange information to absorb more global knowledge. It is different from RNNs, which require complete orders such as left-to-right orders.

C. Graph Attention Networks (GATs)

The graph attentional layer constitutes a fundamental feature of Graph Attention Networks (GATs). They were first described in work by Petar Velickovic et al. [5]. The attention mechanism is applied in a shared manner to all edges in the graph. Therefore it does not depend on upfront access to the global graph structure or features of all of its nodes (which is a limitation of many prior techniques). As opposed to GCNs, GATs assign different importance to nodes of the same neighborhood, enabling a leap in model capacity.

III. GNN ARCHITECTURES IN COMPUTER VISION

To analyze the main concepts present in GNNs applications for computer vision, we have conducted an extensive review of existing literature on the subject. The bibliographical query was based on the Scopus database, using citations of the original Scarcelli paper [1] widely recognized as the source of the GNN concept. The result of the query was limited to the keywords "Computer Vision" or "Object Detection." Repetitive studies with structural advances were treated as redundant and removed from the results list, which covers 100 papers. All papers were then categorized. As the domain of computer vision lacks strict taxonomy, we have used

subjective areas of the Conference on Computer Vision and Pattern Recognition (CVPR'20) conference – one of the most prestigious venues for computer vision research. We have covered here only non-empty sub-domains and created two additional ones. "Mesh Generation and Reconstruction" and "Social Relations Detection and Recognition" were added to accommodate papers that would not fit elsewhere. This Section contains a review of the key novel architectural designs contributed within each category.

A. Action and behavior recognition

This area contains GNN application like action recognition, action segmentation [6],skeleton-based action recognition [7] [8] [9] [10] [11] [12] [13] [14] [15] ,action classification [16]. Recurring architectures which we have identified will be discussed in the following subsections.

1) Spatial-Temporal Graph Convolutional Networks (ST-GCN): Spatial-Temporal Graph neural Networks (STGNNs) aim to discover hidden patterns from spatial-time graphs that became important in various applications, such as recognizing human actions. The key idea of STGNNs is to take into account spatial and temporal dependence simultaneously. Current approaches integrate graph convolutions to capture spatial relationships with RNNs or CNNs for temporal modeling. In computer vision, one of the most significant progress in recent years that has drawn increasing attention is in person re-identification. In this case, a still very challenging task is to effectively overcome the problems of occlusion and the ambiguity for visually similar negative samples. On the other hand, different frames of a video can provide complementary information for each other, and the structural information of pedestrians can provide extra discriminative cues for appearance features. So, building upon the general concept of STGNNs, J.Yang et al. [17] proposed Spatial-Temporal Graph Convolutional Network (ST-GCNs), which includes two convolutional graph layers, where one is spatial and the second is temporal. The spatial part extracts structural information of a human body, while the temporal (discriminative part) mines cues from adjacent frames.

2) Stacked Spatio-Temporal Graph Convolutional Networks: Stacked Spatio-Temporal Graph Convolutional Networks (SPTGCN) was introduced as a solution to leverage the advantages of an encoder-decoder design for improved generalization performance and localization accuracy. Throughout the research, Ghosh et al [6] explored descriptors such as frame-level Visual Geometry Group(VGG), segment-level Inflated 3D ConvNets(I3D), and RCNN-based objects as node descriptors. This research aimed to enable segmentation based on joint inference over the comprehensive contextual one. SPTGCN can be applied in situations that require structured inference over long sequences with heterogeneous data types and various temporal extent.

B. Biometrics, face, gesture, body pose

Biometrics, face, gesture, body pose is one of the broadest categories which usability and approachability increases year

by year. In this category we have identified research issues such as person re-identification [18] [19] [17], human-object interactions (HOI) [20] [21],human pose regression and estimation [22] [23] [24] [25], face recognition [26] , human trajectory prediction [27], group emotion recognition [28], pedestrian detection [29], person search [30], face clustering [31], human parsing [32], group activity recognition [33], classify perceived human emotion from gaits [34], crowd counting and localization [35] and facial action unit intensity [36]. Some architectural advances present in this area are covered below.

1) Graph Mixture Density Network: Graph Mixture Density Network (GraphMDNs) [24] are aimed at combining graph neural networks with mixture density network (MDN). An essential feature of GraphMDNs is their ability to incorporate structured graph information into a neural architecture and the possibility to model multi-modal regression targets. The main distinctive feature of GraphMDNs is the evaluation on random graphs. An additional trait is retained structural information in node representations by computing the distance between distributions of adjacent nodes.

2) Hybrid Graph Neural Network: Hybrid Graph Neural Network (HYGNN) were targeted to relieve the problem with crowd counting were important yet challenging task due to the large scale and density variation by interweaving the multi-scale features. In a paper prepared by Ao Luo et al. [35] they introduce HyGNN as a solution for crowd density because graphs as a structure can improve reasoning process and knowledge representation. HYGNN integrates graphs to jointly represent nodes and two types of relations: multi-scale, capturing the feature dependencies across scales, and mutually beneficial relations building bridges for the cooperation between counting and localization.

3) Masked Graph Attention Networks: Masked Graph Attention Networks (MGATs) [19] operate on features extracted from a graph where nodes are able "to steer" towards other nodes' features under the guidance of label information in the form of a masks matrix. This kind of graph neural network can be used for person re-identification, where it is mainly focused on the correspondence between individual sample images and labels. MGATs can help ignore rich global mutual information residuals in the sample set.

4) Semantic Graph Convolutional Networks: Semantic Graph Convolutional Networks (SemGCNs) [22] were proposed as a solution for the limitation of the small receptive field of convolution filters in GCNs. GCNs in regression tasks are also characterized by a shared transformation matrix for each node in a graph. SemGCNs operate for regression tasks with graph-structured data and learn to capture semantic information to address these limitations. These semantic relations can be learned from the ground truth without additional supervision or are manually integrated into the training process. This kind of graph can be applied to 2D, and 3D human pose regression as introduced and proposed by Zhao et al. in [22].

5) Similarity-Guided Graph Neural Networks: Similarity-Guided Graph Neural Networks (SGGNNs) [18] creates a graph with representation (pair between samples and existing images in several galleries) and utilizes relationships to estimate visual similarities between images better. Unlike GNN approaches, SGGNN learn the edge weights directly with rich labels of gallery instance pairs, and it provides more precise information and alleviates relations' fusion. This happens during the updating process. The Probe-gallery relation feature is then constructed by the messages passing in SGGNNs, taking other nodes for similarity estimation.

6) Spatial-Temporal Graph Attention networks: Spatial-Temporal Graph Attention Network (STGAT) is an architecture built on the sequence-to-sequence scheme but spatial interactions captured by graph attention network (GAT) mechanism. Operation captured spatial is running each time-step. Besides that, it adopts Long-Short Term Memory (LSTM) to encode the temporal correlations of interactions. LSTM is used, e.g., to capture the dynamic interactions of pedestrians, where the latent motions are represented with the hidden states of LSTMs, and GAT is used to aggregate hidden states of LSTMs. This solution was aimed to predict future trajectories of pedestrians and was published in [16], [34].

C. Medical, biological and cell microscopy

Graph Neural Networks are also used in medical diagnosis and the medical image segmentation process. Here we have identified a single study introducing Cell Graph Convolutional Neural Networks, which usability and accuracy can boost and accelerate the detection process of cancer cell detection in medical images.

1) CGC-Net: Cell Graph Convolutional Network: Cell Graph Convolutional Network (CGC-Net) [37] are primarily used for histology image classification. A cell graph is built from an image. In GCC-Net, nuclei are regarded as the nodes and the potential cellular interactions as edges of the graph. In order to obtain a nuclear cell, CGC-Net uses a segmentation network to extract appearance features based on the segmented foreground instances. Cell Graph Convolutional Neural Networks take the whole graph as input to obtain a compact representation of the tissue micro-environment for cancer classification.

D. Computational photography, image and video synthesis

In this subfield we have identified research studies dealing with object reconstruction [38], aesthetics prediction [39] and clustering image [40] issues. Technical advances regarding GNNs include one architecture, covered below.

1) RGNet: The work [41] covers the RGNet architecture. It takes as input an image, representing it as a graph of local regions, and forwards it to a Fully Convolutional Network (FCN) style image aesthetics feature encoder. It is dedicated to producing task-tailored features for aesthetics assessment. The output of the FCN is a 3D feature map in which each spatial location represents one local region in the image. The spatial configuration of all spatial locations in the feature map characterizes the spatial layout of different visual elements in the image. Based on the features of local regions, a region

composition graph is constructed. Each node of this graph is a local region, and any two nodes are connected by an edge-weighted by their feature similarity.

E. Low-level and physics-based vision

Following assumed categorization, we recognize low-level and physics-based vision as a separate subcategory in computer vision. In this subfield we included image restoration [42] and image generation [43] issues. No significant advances to new architectural designs have been provided within the studied references set in this context.

F. Motion and tracking

Motion detection [44] [45] and tracking [46], [47] is another category of used graph inspired neural network in computer vision area. Again, no novelty in architectural designs has been identified within this subfield's contributions.

G. Recognition (object detection, categorization)

Object detection is a key field in artificial intelligence, allowing computer vision to "see" their environment by detecting objects in visual images or videos. This category encompasses object detection [48] [49] [50] [51] [52] [53],image classification [54],image classification [55],visual recognition [56] [57] and recognizing ingredients [58]. New architectures studied in this context were covered below.

1) Cascade Graph Neural Networks: Cascade Graph Neural Networks (Cas-GNN) processes the two data sources individually and employs a novel Cascade Graph Reasoning (CGR) module to learn powerful dense feature embeddings, from which the saliency map can be easily inferred. Modeling and reasoning of high-level relations between complementary data sources explicitly allow us to overcome challenges such as occlusions and ambiguities better.

2) Deep Generative Probabilistic Graph Neural Networks: Deep Generative Probabilistic Graph Neural Networks (DG-PGNN) are used inter alia to generate a scene graph for an image. The input to DG-PGNN is an image with a set of region-grounded captions and object bounding-box proposals. Scene graph generation with DG-PGNN constructs and updates a new model called a Probabilistic Graph Network (PGN). A Probabilistic Graph Network can be thought of as a scene graph with uncertainty. It represents each node and each edge by a CNN feature vector and defines a probability mass function for the node-type of each node and edge-type of each edge. The DG-PGNN sequentially adds a new node to the current PGN by learning the optimal ordering in a Deep Q-learning framework, where states are partial PGNs, actions choose a new node, and rewards are defined based on the ground truth. After adding a node, DG-PGNN uses message passing to update the feature vectors of the current PGN by leveraging contextual relationship information, object co-occurrences, and language priors from captions. The updated features are then used to fine-tune the Probabilistic Matrix Factorization (PMFs) [49].

3) Edge-labeling graph neural networks: Edge-labeling graph neural networks (EGNNs) adapt deep learning networks for tasks with few-shot learning [51]. Neural network-inspired graphs aim to predict the edge labels rather than the node labels on the graphs. Parameters in this network are learned by episodic training with an edge-labeling loss. EGGNs learn to predict edges using intra-cluster and inter-cluster similarities that enable explicit clustering evolution. Learning in EGNNs allows obtaining the generalized model for low-data multiclass problems without retraining.

4) Graph Neural Network Denoising Autoencoders: Denoising Autoencoder network (DAE) takes as input a set of classification weights corrupted with Gaussian noise and learns to reconstruct the target-discriminative classification weights [50]. In this case, the injected noise on the classification weights serves to regularize the weight generating meta-model. DAE model is used as a Graph Neural Network (GNN) to capture the co-dependencies between different classes.

5) LatentGNN: LatentGNN aims to efficiently encode the long-range dependencies instead of relying on a fully-connected graph defined on input features. LatentGNN was introduced as a solution to Learning Efficient Non-local Relations for Visual Recognition. To construct LatentGNN Restricted Boltzmann Machine (RBM) [59] with undirected graph models with latent representations are being used. In LatentGNN, the original feature nodes are augmented with a set of additional latent nodes, and then these nodes are connected structurally.

6) Multi–Relational Graph Convolutional Network: Multi-Relational Graph Convolutional Networks (mRGCN) were proposed as a solution to integrate ingredient hierarchy attribute for zero-shot ingredient recognition [58]. Recognition with mRGCN can develop models that can identify unseen ingredients, as obtaining a fully annotated training set is sometimes too complex due to the high visual variance of input data. mRGCN comprises two modules: mRCGN works for the unseen ingredient prediction, and Deep Convolutional Neural Networks Classifier deals with already known ingredients. Then the obtained trained classifier is used to generate ground-truth supervising the learning process of mRGCN.

7) OD-GCN: Object Detection Boosted GCN: OD-GCN (Object Detection with Graph Convolutional Networks) was proposed to boost the detection performance. The architecture of OD-GCN combines two models: Object Detection (OD) and Graph Convolutional Network (GCN). OD model represents any classical object detection model, so in OD-GCN we can use, e.g., Fast-RNN or others. The second part of OD-GCN is a model based on knowledge graphs that can be used for post-processing.

8) Residual Attention Graph Convolutional Networks: Residual Attention Graph Convolutional Networks (RAGCNs) constitute graph-based variants of residual neural networks using attention mechanisms and residual solutions to exploit the intrinsic geometric context inside a 3D space. RAGCNs [52] allow the use of organized or unorganized 3D data without

using any point features, which increases the effectiveness of the proposed architecture.

H. Scene Analysis and Understanding

Scene graph is a data structure, which is used for scene generation [60]–[63]. Sometimes we want to know more about objects relationships on a scene [64]. Graphs neural networks seem to be a natural choice for generating such scene graphs. We have identified one architecture advance within this subfield.

1) Dynamic Gated Graph Neural Network: Dynamic Gated Graph Neural Network (D-GGNN) [62] is a deep generative architecture that can be applied to an input image when only partial relationship information, or none at all, is known. It is an upgrade to the Gated Graph Neural Networks (GGNN) [65]. This latter network type cannot work when input data is partial, which is a severe limitation in computer vision where missing data often occurs. D-GGNN, similarly to Generative Probabilistic Graph Neural Networks (DG-PGNN) are frequently used for scene graph generation tasks. This architecture builds the graph in the process of reinforcement learning (RL). The graph built so far in each training step is the current state. The D-GGNN leverages the power of GGNN to exploit the relational context information of the input image and combines the gated graph neural network with an attention mechanism.

I. Segmentation, grouping, and shape

In this area, we keep focusing on usability and the possibility to apply Graph Neural Networks in segmentation, grouping, and shape area. In this subfield, we have identified one novel GNN structure aimed at addressing the problem of referring expression comprehension in videos [66].

1) Language-Guided Graph Attention Networks: Language-Guided Graph Attention Networks (LGRANs) highlight the relevant content referred to by the expression. The graph attention comprises two main components: a node attention component to highlight relevant objects and an edge attention component to identify the object relationships present in the expression. In LGRANs additionally, encode the relationships between objects (that have properties of their own). Secondly, in LGRANs, attention is guided by the referring statement. Thirdly, LGRANs, unlike GATs, maintain different types of features to represent the node properties and node relationships.

J. Video analysis and understanding

In this subfield we have identified the application of GNN in Video understanding models [67] or video anomaly detection under weak labels [68]. Three specific architectural types are covered below.

1) Curve-GCN: Curve-GCN is used for interactive object annotation, using parametrized objects with either polygons or splines, and is trained end-to-end at a high output resolution. Curve-GCN alleviates the sequential nature of PolygonRNN, by predicting all vertices simultaneously using a Graph Convolutional Network (GCN) [69].

2) Social Relationship Graph Generation Networks: Social Relationship Graph Generation Network (SRG-GN) proposed by Goel et al [70] uses memory cells like Gated Recurrent Units (GRUs) to iteratively update the social relationship states in a graph using scene and attribute context. SRG-GN can unify the representation of social relationships and attributes forming input images. This architecture is based on recurrent connections among the GRUs. They were implemented for passing messages between nodes and edges, improving GRUs social relationship recognition.

K. Vision applications and systems, vision for robotics and autonomous vehicles

It is one of the two most extensive areas of computer vision. Research in this category is more application-oriented and focuses on solutions for available problems rather than architectural advances and preparing complex solutions. In this category, we can find research studies dealing with modeling edge features with Deep Bayesian Graph Networks [71], Graph Convolutional Gaussian Processes [72], GCAN Graph Convolutional Adversarial Network [73], data representation and learning with Graph Diffusion-Embedding Networks [74], extracting building polygons [75] or temporal prediction [76]. Three architectural advances in this subfield will be named below.

1) Graph Convolutional Adversarial Networks: Graph Convolutional Adversarial Networks (GCANs) were proposed for unsupervised domain adaptation. GCAN [73] is a model which uses three kinds of inputs: structure, domain label, and class label to generate the deep model for unsupervised domain adaptation. GCANs employ structure-aware alignment, domain alignment, and class centroid alignment to reduce the domain discrepancy for domain adaptation.

2) Graph Diffusion-Embedding Networks: Graph Diffusion-Embedding Networks (GDENs) were proposed as a method of graph-based feature diffusion. GDENs [74] integrate both feature diffusion and graph node embedding simultaneously into a unified network. By employing a novel naturally scalable diffusion-embedding architecture with diffusion-embedding operation, a simple closed-form solution can be obtained, which thus guarantees high efficiency.

3) Graph Convolutional Gaussian Processes: Graph Convolutional Gaussian Processes [72] represent a novel Bayesian nonparametric method in non-Euclidean domains. Graph Convolutional Gaussian Processes can be applied to problems in machine learning, for which the input observations are functions with domains on general graphs.

L. Vision + other modalities, Vision + language

Second largest of category of CVPR in terms of number of contributions is a Vision + other modalities, Vision + language. In this category we have identified papers on image annotation [70], [77]–[79], visual question answering [80]–[85], situation recognition [86], [87], as well as image understanding, image captioning and object annotation [69]. In this subfield no novelty in architectural designs have been identified within GNN-based contributions.

Dataset name	Fields of application	Details
NTU-RGBD	Action and behavior recognition	60 action classes and 56,880 video samples
CUHK03	Biometrics, face, gesture, body pose	14,097 images
Skeleton-Kinetics	Action and behavior recognition	300,000 videos
Visual Genome	Scene analysis	108,077 images
Visual Genome	Low-level and physics-based vision	108,077 images
Visual Genome	Mesh Generation and reconstruction	108,077 images
ShapeNet	Mesh Generation and reconstruction	55 common object categories
ShanghaiTech	Video analysis and understanding	1198 annotated crowd images
COCO	Recognition (object detection, categorization)	over 320,000 images
CoNSeP	Medical, biological and cell microscopy	41 HE stained image tiles
Colorectal Cancer (CRC)	Medical, biological and cell microscopy	139 images
MS-Caleb-1M	Computational photography, image and video synthesis	10 million face images
Jilin-1 and SkySat	Motion and tracking	8 satellite video sequences

M. Mesh Generation and Reconstruction

This category was introduced to accommodate studies in mesh generation [88]–[91] and reconstruction [92] in which graph inspired neural networks can be naturally employed.

IV. DATASETS

Contributions presented in this paper cover a variety of applications which require for their validation benchmarks of different nature. Still, there exist a common set of benchmarks which are omnipresent in the GNN-based computer vision studies. The list of most used datasets, along with their field of appliactions and characteristics, is presented in Table I.

V. RELATIONSHIPS AND INSPIRATIONS

To visualize relations between manuscripts described in this paper, we used a graph-based approach. It was employed to show relations between papers we have covered in our study. Relationships are represented by edges and publications by nodes. Marker size corresponds to the number of times the paper has been cited in the other articles from the analyzed set. Connections between articles and references are based on search engine and API of Semantic Scholar which indexes publications published by major scientific publishers (including IEEE, ACM, Springer, or Wiley). The resulting graph is presented on Figure 2.

The resulting graph is presented on Figure 2. It can be seen that clusters of articles, similar to the groups of CVPR conference topics, can be identified. It allows us to distinguish papers from similar methodological backgrounds, though often not sharing the same scope or architecture. We can also identify papers present in the graph (e.g. [55], [91]) which have significant impact on publications from another clusters. This also concerns the major publications in another clusters (like [67] has impact on [91]).

On Figures 3 and on 4 we have demonstrated how significant impact for the analyzed publications – measured by citation numbers – have other articles, not only in the domain of computer vision. The larger the size of the marker of a given article, the more significant the impact. We can see how big impact have on computer vision based on graph neural network have fundamental contributions [1], [3] prepared by F.Scarselli et al and by Kipf et al. Both graphs' source codes are made available in the repository to allow their replication.

In Table II we have presented the top 20 most cited works within studied articles. After analysis of these references,
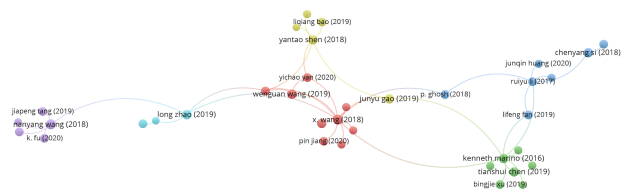

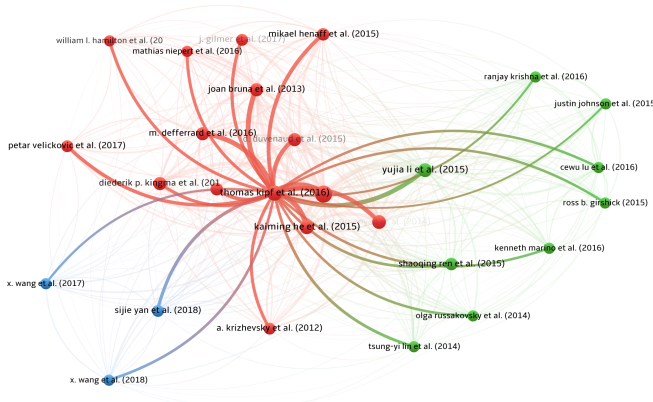Fig. 2. Visualization of links between GNN-based computer vision papers

Fig. 3. Visualization of the impact of the Kipf et al paper "Semi-Supervised Classification with Graph Convolutional Networks in computer vision"
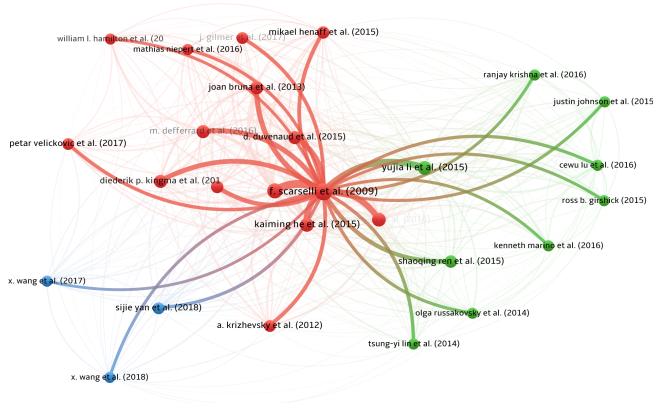

Fig. 4. Visualization of the impact of the Scarselli et al paper "The Graph Neural Network Model in computer vision"

we can see that in addition to the base article in the field of graph neural networks [1], issues related to, inter alia, with recursive networks [107] [106] [96] had a significant impact on the development of Graph Neural Networks in the area of computer vision. Moreover, we can see that the graph's attention mechanism [5] and convolutional operations on graphs [3] had a significant influence on the development of Graph Neural Networks-based methods in this field.

VI. CONCLUSION

In the last few years, studies on GNNs have become an active research field, mainly due to the advance of underlying Deep Learning architectures. The essential idea of GNNs is to enhance the node representations by propagating information between nodes. This paper provided an analysis of state-of-art Graph Neural Networks in computer vision. We delivered a thorough review of the latest achievements, summarizing architectural advances. We also studied the relations between methods and identified sources of inspiration and related studies, also outside of the explored domain. We have discovered that studies on skeleton-based action recognition and research in one subfield – biometrics, face, gesture, body pose – seem to be very influential for developing vision systems using GNNs. We anticipate that the results of this study might be useful for researchers, not only in the field of computer vision.

ACKNOWLEDGMENTS

REFERENCES

[1] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[2] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *ArXiv*, vol. abs/1506.05163, 2015.

[3] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2017.

[4] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 764–10 773.

[5] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv*, 2018.

[6] P. Ghosh, Y. Yao, L. S. Davis, and A. Divakaran, "Stacked spatio-temporal graph convolutional networks for action segmentation," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 565–574, 2020.

[7] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 7904–7913.

[8] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Part-level graph convolutional network for skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11 045–11 052, Apr. 2020.

[9] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Adaptive spectral graph convolutional networks for skeleton-based action recognition," *CoRR*, vol. abs/1805.07694, 2018. [Online]. Available: http://arxiv.org/abs/1805.07694

[10] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," *CoRR*, vol. abs/1902.09130, 2019. [Online]. Available: http://arxiv.org/abs/1902.09130

[11] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," *CoRR*, vol. abs/1904.12659, 2019. [Online]. Available: http://arxiv.org/abs/1904.12659

[12] Y. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. hong Xia, "Graph cnns with motif and variable temporal block for skeleton-based action recognition," in *AAAI*, 2019.

[13] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer, 2018, pp. 106–121.

[14] J. Huang, Z. Huang, X. Xiang, X. Gong, and B. Zhang, "Long-short graph memory network for skeleton-based action recognition," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 634–641.

[15] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition

by neural searching," *CoRR*, vol. abs/1911.04131, 2019. [Online]. Available: http://arxiv.org/abs/1911.04131

[16] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *CoRR*, vol. abs/1801.07455, 2018. [Online]. Available: http://arxiv.org/abs/1801.07455

[17] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao, "Learning multi-granular hypergraphs for video-based person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.

[18] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Computer Vision – ECCV 2018*. Springer, 2018, pp. 508–526.

[19] L. Bao, B. Ma, H. Chang, and X. Chen, "Masked graph attention network for person re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Jun. 2019.

[20] S. Qi, W. Wang, B. Jia, J. Shen, and S. Zhu, "Learning human-object interactions by graph parsing neural networks," *CoRR*, vol. abs/1808.07962, 2018. [Online]. Available: http://arxiv.org/abs/1808.07962

[21] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[22] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[23] H. Ci, C. Wang, X. Ma, and Y. Wang, "Optimizing network structure for 3d human pose estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[24] T. Oikarinen, D. Hannah, and S. Kazerounian, "GraphMDN: Leveraging graph structure and deep learning to solve inverse problems," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2021.

[25] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo, "Differentiable hierarchical graph grouping for multi-person pose estimation," in *Computer Vision – ECCV 2020*. Springer, 2020, pp. 718–734.

[26] J. Zheng, R. Yu, J.-C. Chen, B. Lu, C. Castillo, and R. Chellappa, "Uncertainty modeling of contextual-connections between tracklets for unconstrained video-based face recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[27] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial-temporal interactions for human trajectory prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[28] X. Guo, L. F. Polania, B. Zhu, C. Boncelet, and K. E. Barner, "Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2020.

[29] G. Li, J. Li, S. Zhang, and J. Yang, "Learning hierarchical graph for occluded pedestrian detection," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, Oct. 2020.

[30] Y. Yang, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," *CoRR*, vol. abs/1904.01830, 2019. [Online]. Available: http://arxiv.org/abs/1904.01830

[31] L. Yang, X. Zhan, D. Chen, J. Yan, C. C. Loy, and D. Lin, "Learning to cluster faces on an affinity graph," *CoRR*, vol. abs/1904.02749, 2019. [Online]. Available: http://arxiv.org/abs/1904.02749

[32] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," *CoRR*, vol. abs/1904.04536, 2019. [Online]. Available: http://arxiv.org/abs/1904.04536

[33] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," *CoRR*, vol. abs/1904.10117, 2019. [Online]. Available: http://arxiv.org/abs/1904.10117

[34] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "STEP: spatial temporal graph convolutional networks for emotion perception from gaits," *CoRR*, vol. abs/1910.12906, 2019. [Online]. Available: http://arxiv.org/abs/1910.12906

[35] A. Luo, F. Yang, X. Li, D. Nie, Z. Jiao, S. Zhou, and H. Cheng, "Hybrid graph neural networks for crowd counting," *CoRR*, vol. abs/2002.00092, 2020. [Online]. Available: https://arxiv.org/abs/2002.00092

[36] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution," *CoRR*, vol. abs/2004.09681, 2020. [Online]. Available: https://arxiv.org/abs/2004.09681

[37] Y. Zhou, S. Graham, N. Alemi Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot, "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 388–398.

[38] K. Fu, J. Peng, Q. He, and H. Zhang, "Single image 3d object reconstruction based on deep learning: A review," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 463–498, Sep. 2020.

[39] D. Liu, R. Puri, N. Kamath, and S. Bhattacharya, "Composition-aware image aesthetics assessment," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 3558–3567.

[40] Z. Wang, L. Zheng, Y. Li, and S. Wang, "Linkage based face clustering via graph convolution network," *CoRR*, vol. abs/1903.11306, 2019. [Online]. Available: http://arxiv.org/abs/1903.11306

[41] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[42] J. Qiu, Y. Yang, X. Wang, and D. Tao, "Hallucinating visual instances in total absentia," in *Computer Vision – ECCV 2020*. Springer, 2020, pp. 264–282.

[43] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228, 2018.

[44] J. Feng, D. Zeng, X. Jia, X. Zhang, J. Li, Y. Liang, and L. Jiao, "Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 116–130, 2021.

[45] K. Hu, Z. Wang, S. Mei, K. A. E. Martens, T. Yao, S. J. G. Lewis, and D. D. Feng, "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 1215–1225, Apr. 2020.

[46] R. Yeh, A. Schwing, J. Huang, and K. Murphy, "Diverse generation for multi-agent sports games," in *Proc. CVPR*, 06 2019, pp. 4605–4614.

[47] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4644–4654, 2019.

[48] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, "Cascade graph neural networks for RGB-d salient object detection," in *Computer Vision – ECCV 2020*. Springer, 2020, pp. 346–364.

[49] M. Khademi and O. Schulte, "Deep generative probabilistic graph neural networks for scene graph generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11 237–11 245, Apr. 2020.

[50] S. Gidaris and N. Komodakis, "Generating classification weights with GNN denoising autoencoders for few-shot learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[51] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[52] A. Mosella-Montoro and J. Ruiz-Hidalgo, "Residual attention graph convolutional network for geometric 3d scene classification," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, Oct. 2019.

[53] Z. Liu, Z. Jiang, W. Feng, and H. Feng, "OD-GCN: Object detection boosted by knowledge GCN," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, Jul. 2020.

[54] W. S. Campos, L. G. Souza, P. T. M. Saito, and P. H. Bugatti, "Contextual image classification through fine-tuned graph neural networks," in *Artificial Intelligence and Soft Computing*. Springer, 2021, pp. 15–24.

[55] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017.

[56] Y. Li and A. K. Gupta, "Beyond grids: Learning graph representations for visual recognition," in *NeurIPS*, 2018.

[57] S. Zhang, S. Yan, and X. He, "Latentgnn: Learning efficient non-local relations for visual recognition," *CoRR*, vol. abs/1905.11634, 2019. [Online]. Available: http://arxiv.org/abs/1905.11634

[58] J. Chen, L. Pan, Z. Wei, X. Wang, C.-W. Ngo, and T.-S. Chua, "Zero-shot ingredient recognition by multi-relational graph convolutional network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10 542–10 550, Apr. 2020.

[59] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines vinod nair," in *Proceedings of ICML*, vol. 27, 06 2010, pp. 807–814.

[60] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4560–4568, 2019.

[61] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6156–6164.

[62] M. Khademi and O. Schulte, "Dynamic gated graph neural networks for scene graph generation," in *Computer Vision – ACCV 2018*, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Springer, 2019, pp. 669–685.

[63] A. Dornadula, A. Narcomey, R. Krishna, M. Bernstein, and F.-F. Li, "Visual relationships as functions:enabling few-shot scene graph prediction," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1730–1739.

[64] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined cnn and gcn for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4325–4338, 2020.

[65] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arxiv:1511.05493*, 2015. [Online]. Available: http://arxiv.org/abs/1511.05493v4

[66] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. van den Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," *CoRR*, vol. abs/1812.04794, 2018. [Online]. Available: http://arxiv.org/abs/1812.04794

[67] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer, 2018, pp. 413–431.

[68] J. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," *CoRR*, vol. abs/1903.07256, 2019. [Online]. Available: http://arxiv.org/abs/1903.07256

[69] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, "Fast interactive object annotation with curve-gcn," *CoRR*, vol. abs/1903.06874, 2019. [Online]. Available: http://arxiv.org/abs/1903.06874

[70] A. Goel, K. T. Ma, and C. Tan, "An end-to-end network for generating social relationship graphs," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[71] D. Atzeni, D. Bacciu, F. Errica, and A. Micheli, "Modeling edge features with deep bayesian graph networks," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2021.

[72] I. Walker and B. Glocker, "Graph convolutional gaussian processes," *CoRR*, vol. abs/1905.05739, 2019. [Online]. Available: http://arxiv.org/abs/1905.05739

[73] X. Ma, T. Zhang, and C. Xu, "GCAN: Graph convolutional adversarial network for unsupervised domain adaptation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[74] B. Jiang, D. Lin, J. Tang, and B. Luo, "Data representation and learning with graph diffusion-embedding networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[75] S. Wei and S. Ji, "Graph convolutional networks for the automated production of building vector maps from aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[76] I. Kostrikov, Z. Jiang, D. Panozzo, D. Zorin, and J. Bruna, "Surface networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018.

[77] J. Zhang, Q. Wu, J. Zhang, C. Shen, and J. Lu, "Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[78] A. Kumar, A. A. Shenoy, and A. Sharma, "MKL based local label diffusion for automatic image annotation," in *Communications in Computer and Information Science*. Springer, 2018, pp. 385–399.

[79] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-RNN," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018.

[80] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," *arxiv:1511.02799*, 2015. [Online]. Available: http://arxiv.org/abs/1511.02799v4

[81] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," *arxiv:1609.05600*, 2016. [Online]. Available: http://arxiv.org/abs/1609.05600v2

[82] W. Norcliffe-Brown, E. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," *arxiv:1806.07243*, 2018. [Online]. Available: http://arxiv.org/abs/1806.07243v6

[83] M. Narasimhan, S. Lazebnik, and A. G. Schwing, "Out of the box: Reasoning with graph convolution nets for factual visual question answering," *arxiv:1811.00538*, 2018. [Online]. Available: http://arxiv.org/abs/1811.00538v1

[84] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," *arxiv:2008.09105*, 2020. [Online]. Available: http://arxiv.org/abs/2008.09105v1

[85] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11 109–11 116, apr 2020.

[86] M. Suhail and L. Sigal, "Mixture-kernel graph attention network for situation recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[87] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, "Situation recognition with graph neural networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[88] R. Daroya, R. Atienza, and R. Cajote, "REIN: Flexible mesh generation from point clouds," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Jun. 2020.

[89] A. Dai and M. NieBner, "Scan2mesh: From unstructured range scans to 3d meshes," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[90] J. Tang, X. Han, J. Pan, K. Jia, and X. Tong, "A skeleton-bridged deep learning approach for generating meshes of complex topologies from single RGB images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[91] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single RGB images," in *Computer Vision – ECCV 2018*. Springer, 2018, pp. 55–71.

[92] E. J. Smith, S. Fujimoto, A. Romero, and D. Meger, "Geometrics: Exploiting geometric structure for graph-encoded objects," *CoRR*, vol. abs/1901.11461, 2019. [Online]. Available: http://arxiv.org/abs/1901.11461

[93] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[94] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[95] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3844–3852.

[96] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[97] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[98] D. K. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. D. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," *ArXiv*, vol. abs/1509.09292, 2015.

[99] J. Bruna, W. Zaremba, A. D. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *CoRR*, vol. abs/1312.6203, 2014.

[100] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2, pp. 729–734 vol. 2, 2005.

[101] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[102] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *ArXiv*, vol. abs/1704.01212, 2017.

[103] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[104] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017.

[105] X. Wang, R. B. Girshick, A. K. Gupta, and K. He, "Non-local neural networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[106] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[107] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 386–397, 2020.