視線情報を埋め込んだテンポラルシーングラフを用いたナビゲーション文章生成による Human-like Guidance

難波田 雅己† 宮田 昌樹† 大久保 翔太‡ 南里 卓‡ 平川 翼† 山下 隆義† 藤吉 弘亘†
†中部大学 ‡日産自動車株式会社
E-mail: masaknanbt@mprg.cs.chubu.ac.jp

1 はじめに

自動車の案内として，機械によるカーナビゲーションと人間によるナビゲーションの 2 つが考えられる．カーナビゲーションとは既存のカーナビゲーションシステムのことであり，車載モニターのデジタル地図と，右左折地点からの距離や方角，道路名を基準にした案内文によりナビゲーションをする．そのため,「100m 先を右方向です」「〇〇通りを北に進んでください」といった定型文による案内であるが，直感的な理解が難しいことがある．よって，複雑な構造の道の場合ナビゲーションを誤解してしまうことや，モニターを注視して注意力が散漫になってしまう等の問題がある．一方，人間によるナビゲーションでは，同乗者が視認した周辺環境の情報を利用して，建物や周辺車等を利用して案内を行う．そのため,「ガソリンスタンドがある交差点を右方向です」「前の黄色い車に付いて行ってください」と言った状況に合わせた案内になり，直感的に理解しやすい．また，人間は空間を認知する際に距離ではなく建物等の物体を用いる傾向がある [1]．よって，カーナビゲーションによる問題を防ぐことが可能であると考えられるが，同乗者がいなければ成り立たない．
これらの問題点を解決するために，本研究では人間のナビゲーションをシステムで再現することを目標とした Human-like Guidance (HℓG) の実現を目的とする．HℓG では,「前方の赤い車を追従して次の交差点を右折です」など状況に合わせたわかりやすい案内文でのナビゲーションがシステムのみで可能となる．本研究では，人間の運転中の視線情報と時系列情報を考慮したテンポラルシーングラフによる HℓG の実現を提案する．人間の運転中の視線情報は，視線推定ネットワークを用いて運転中の視線情報を学習し，運転者がどこを注視しているか推定することで再現する．また，時系列を考慮した周囲の環境情報を表すシーングラフであるテンポラルシーングラフを提案し，推定した視線情報を付与することで，人間が着目しているものをベースとした案内文を生成する．

2 関連研究

本章では，視線推定と，グラフ構造から特徴抽出する手法について述べる．

2.1 視線推定

視線推定とは，人間が画像や動画像を与えられたとき，どこを見ているのかを予測するタスクである．視線推定の代表的な手法の 1 つとして，Dilated Inception Network (DINet) [2] がある．DINet は，入力画像の特徴量から視線を推定し，人間の視線と推定結果を比較することで視線を学習するエンコーダデコーダネットワークである．DINet では，Dilated Residual Network (DRN) [3] をエンコーダとし，Dilated Inception Module (DIM) でマルチスケール特徴を補捉した後，デコーダで顕著性マップを生成する．

DRN は，畳み込み層を 50 層とした Residual Network (ResNet) [4] である ResNet-50 に，拡張畳み込み処理を導入したネットワークである．DRN は，視線推定に用いられることが多い事前学習済みネットワークであり，同じく視線推定に用いられることの多い VGG-16 と比較して，広い受容野の特徴マップを獲得することが出来る．また，計算量も VGG-16 より少ない．

DIM は，GoogLeNet[5] で提案された Inception モジュールに拡張畳み込み処理を導入したものである．Inception モジュールのそれぞれの畳み込み処理を拡張率の異なる拡張畳み込み処理に置き換えることにより，少ないパラメータ数，短い学習時間で同程度またはそれ以上の精度が達成可能である．

2.2 グラフ構造から特徴抽出

グラフ構造から特徴抽出する手法は一般に，Spatial methods と Spectral methods の 2 つに大別される．Spatial methods はユーグリッド空間における演算のみでグラフ特徴を抽出する手法である．代表的な手法として，最も基本的な枠組みを作成した手法である Graph Neural Network (GNN) [6] や，その発展形である Gated Graph Sequence Neural Networks (GGNN) [7] がある．一方，Spectral methods は，グラフに対してフーリエ空間内で畳み込み処理を定義し，特徴を抽出する手法

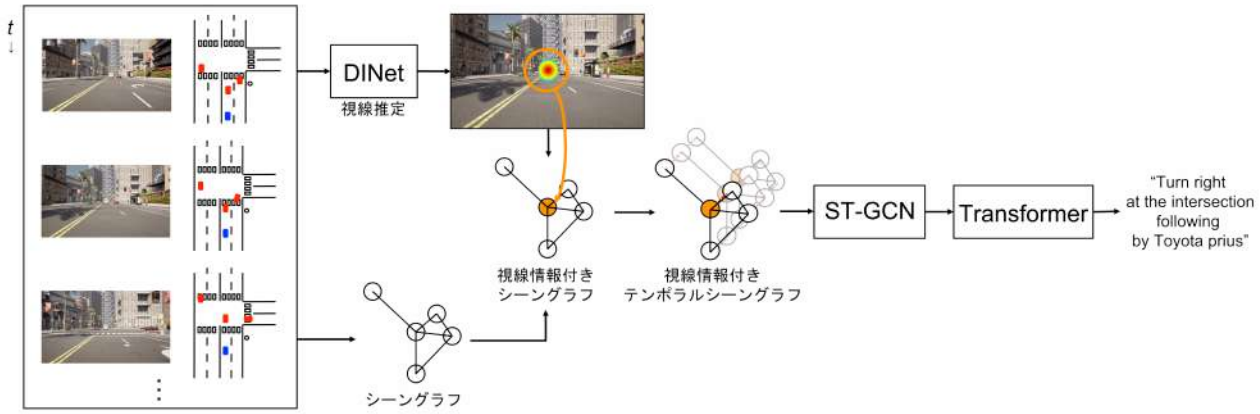


図 1　提案手法の流れ

図 2　シーングラフの作成

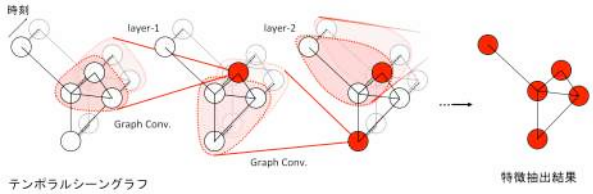


図 3　ST-GCN によるテンポラルシーングラフの特徴抽出

である．代表的な手法として，Bruna ら [8] による，畳み込みの定理を応用したグラフにおける畳み込み処理の定義や，Defferrard ら [9] による，近似式を用いることでグラフ構造の局所性を考慮する手法がある．また近年，Kipf と Welling [10] によって Spatial methods と Spectral methods を統合するような手法が提案された．この手法は，Spectral methods からの系譜を継いだ手法ではあるが，ユークリッド空間での行列演算を用いることで，グラフ特徴を抽出することを可能とした．この手法のみを指して Graph Convolutional Networks (GCN) と呼ぶことも多く，本研究ではこれに沿う．

3　提案手法

本研究では，時系列を考慮した周囲の環境情報を表すテンポラルシーングラフを提案し，視線情報とテンポラルシーングラフによる HℓG の実現手法について述べる．提案手法の概略を図 1 に示す．提案手法ではまず，周囲の状況をシーングラフとして表現する．同時に視線推定モデルを用いて，着目対象を推定し，この視線情報をシーングラフに追加する．次に，視線情報付きシーングラフを任意の時刻分連結した，テンポラルシーングラフを作成する．作成したテンポラルシーングラフを Spatial Temporal Graph Convolutional Network に入力し，テンポラルシーングラフの特徴量を獲得する．最後に，得られた特徴量から，Transformer によって案内文を生成する．

3.1　運転者の視線の推定

人間の運転中の視線を再現するために，視線推定モデルである DINet[2] を用いて視線推定を行う．DINet は RGB 画像を入力とし，視線推定結果をヒートマップとして出力する．ヒートマップから注視物体特定するために，閾値以上の領域の外接矩形を求める．学習では，出力したヒートマップと真値の視線座標を元にしたマップとの誤差を求める．誤差関数には，Sheng ら [2] が提案した線形正規化を用いた関数を使用する．ここで，視線情報には，通常運転時の視線情報ではなく，ナビゲーションを行う際に着目しやすい対称に意図的に誘導を行ったデータを用いる．これによって，HℓG のナビゲーションに適した視線の推定が可能となる．

3.2　視線情報を付与したシーングラフ

時刻 t の周囲の状況のシーングラフ G を，式 (3.2) のように定義する．ノードは $v_i = \{$ オブジェクトクラス，座標，車両名，色 $\}$ のように定義され，エッジにはノード間の相対距離を用いる．

$$
\begin{aligned}
V_i &= \{v_i j i = 1, ..., N\} \\
E_s &= \{v_i, v_j j(i,j) = H\} \\
G &= \{V_i, E_s\}
\end{aligned}
\tag{1}
$$

ここで，N は全体のノード数，H は全体のエッジ数を表す．加えて，視線推定により得られた矩形と位置が

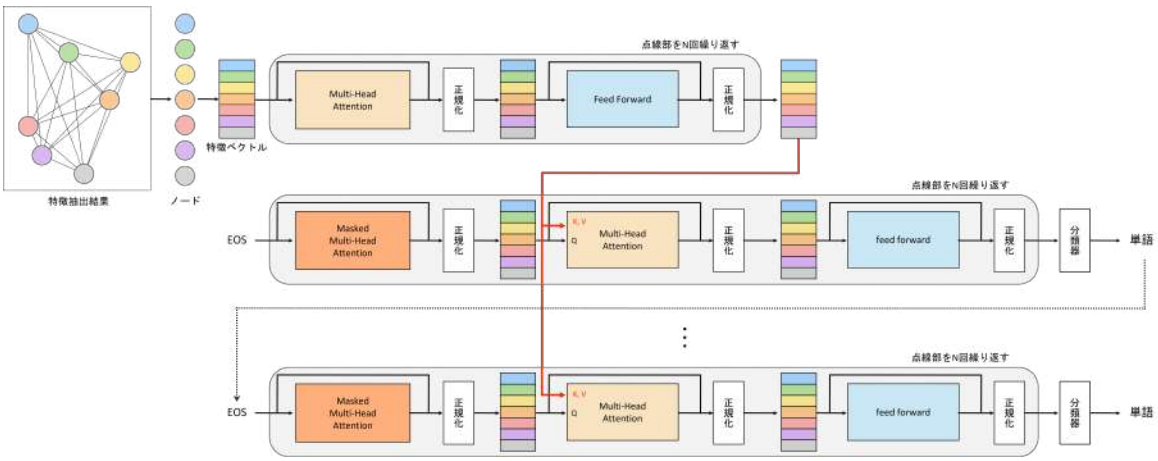


図 4　Transformer による文章生成

一致するオブジェクトを着目対象とする．よって，ノード集合を $vg_i = \{$ オブジェクトクラス，座標，車両名，色，　視線情報の有無 $\}$ と定義する．図 2 に示すように，視線情報を付与したシーングラフを式 (2) のように定義する．

$$G = \{Vg_i, E_s\} \tag{2}$$

3.3　ST-GCN による特徴抽出

各オブジェクトの位置関係について時系列情報を考慮するために，ST-GCN [11] を用いる．現在から未来 T 時刻分について，各時刻のシーングラフのノード集合を $Vg_{ti} = \{vg_{ti} jt = 1, ..., T, i = 1, ..., N\}$，エッジ集合を $E_s = \{v_{ti}, v_{tj} j(i, j) = H\}$ と定義する．また，隣接した他時刻における同じオブジェクトを接続するエッジ $E_f = \{v_{ti}, v_{(t+1)i}\}$ を追加することで，テンポラルシーングラフ G_t を式 (3) のように定義する．

$$G_t = \{Vg_{ti}, E_s, E_f\} \tag{3}$$

図 3 に示すように，グラフ畳み込み処理の際に，各時刻で共通するノードを同時に畳み込む．これにより，オブジェクト毎に移動を考慮した特徴量を獲得する．本研究に用いる ST-GCN は，6 層のグラフ畳み込み層を用いてテンポラルシーングラフから特徴抽出を行う．

3.4　Transformer による案内文生成

図 4 に示すように，ST-GCN で得られたグラフ特徴量から，着目対象を基にした案内文の生成を行う．文章生成のモデルとして Transformer[12] を用いる．Encorder6 層，Decorder6 層で構成されており，Trandformer Encorder では，ST-GCN で得られた各ノードの特徴量をトークンとして入力する．Transformer Decoder では，Encorder から獲得したトークンを Decorder の入力として用い，単語を出力する．この時，教師データとして該当シーンにおける正解となる案内文を用いる．正解の案内文は，該当シーンにおける真値の視線情報を用いてルールベースで作成する．これにより，各オブジェクトに対応したナビゲーション文を生成できる．

Turn right at the intersection, following the toyota prius.

(a) 自車両と進行方向が同じ場合

Turn right at the intersection where the tesla amblance is located.

(b) 自車両と進行方向が異なる場合

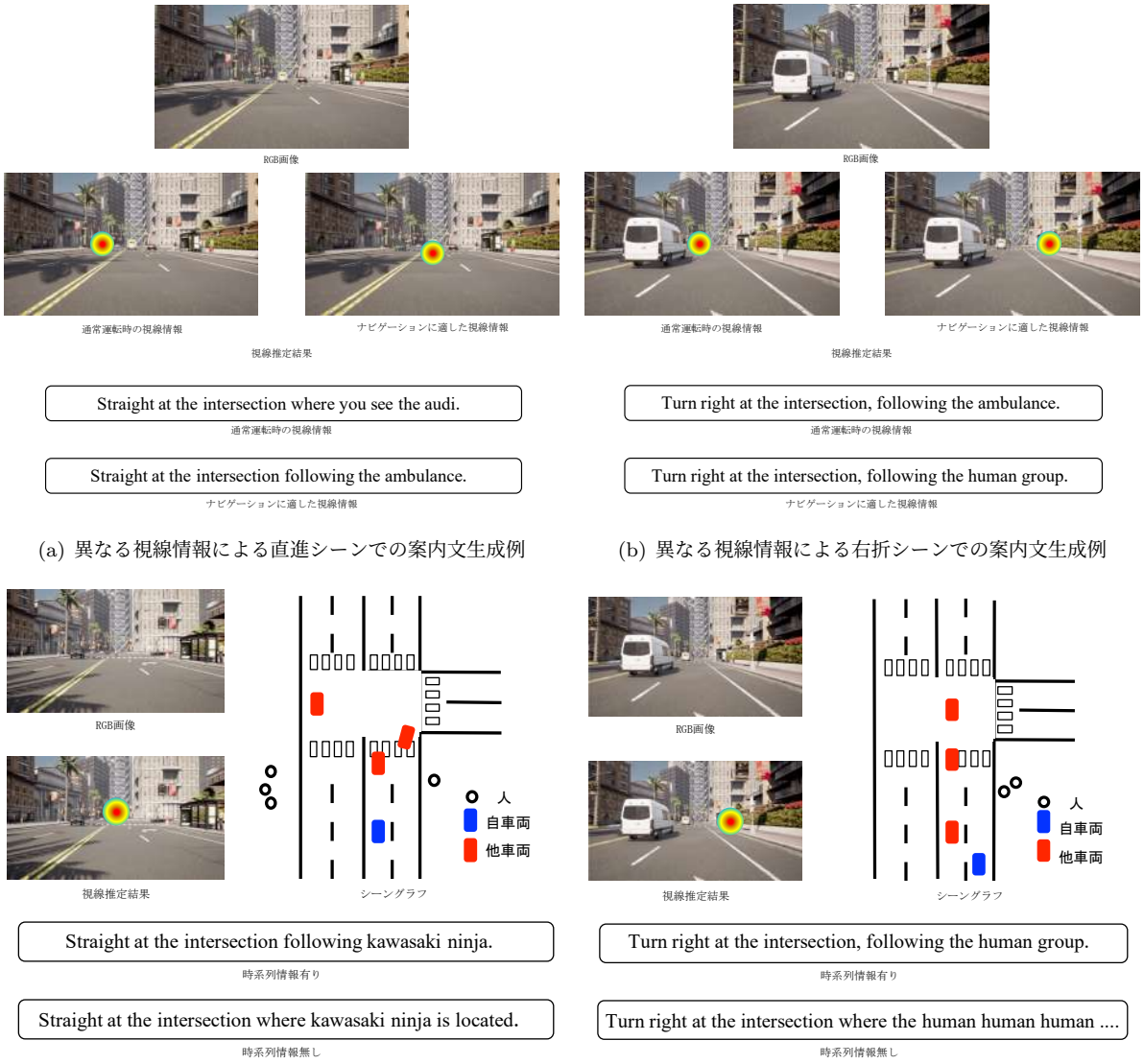
図 5　データセット内の案内文例

4　評価実験

本実験では，視線推定モデルの学習において，意図的な誘導を行った視線データの有効性の検証，提案した手法における時系列情報の有無による影響を，それぞれの実験で生成された文章を用いて評価する．

4.1　データセット

DINet とテンポラルシーングラフのデータセットには，それぞれ自動運転開発シミュレータである CARLA Simulater で作成した動画像を用いる．DINet では，全 20 シーン，合計約 20 万フレームの運転シーンの動画



(a) 異なる視線情報による直進シーンでの案内文生成例

(b) 異なる視線情報による右折シーンでの案内文生成例

(c) 前方車追従シーンの案内文生成例

(d) 右折シーンの案内文生成例

図 6　各実験における案内文比較

像と，それらに対する人間の視線情報をデータセットとする．テンポラルシーングラフからの文書生成では，全 14 シーン，合計 1672 フレームの運転シーンの動画像と，各フレームに存在する全てのオブジェクトに対する案内文をデータセットとする．ここで，案内文はルールベースでオブジェクトに対する文章である．オブジェクトが人の場合，2 名以上が近距離に位置している時は集団とし，集団に対する案内文とした．案内文作成のルールは，進行方向，対象オブジェクトの名前，進行方向に対する自車両の行動を含む文とすることである．対象オブジェクトは，該当シーンにおける視線情報を用いて選択する．自車両の行動に関する指示は，対象オブジェクトの進行方向と自車量の進行方向が同じであれば追従，異なる場合は対象オブジェクトの位置を用いる．案内文の例を図 5 に示す．

4.2　実験概要

テンポラルシーングラフを作成する際に考慮する時系列情報は，現在時刻から未来 10 時刻分とする．視線推定モデルの有効性を確認する評価では，2 つの視線データの収集方法を比較する．1 つ目は通常運転時の視線データであり，カーナビゲーションのように指示を行って収集した．2 つ目はナビゲーションに適した視線データであり，ナビゲーションを行う際に着目しやすい対象を継続的に見るように指示をして収集した．また，文書生成には，時系列情報ありのモデルを使用する．次に，文書生成の評価として，時系列情報の有無による比較を行う．時系列情報を用いない場合，入力するシーングラフは現在時刻のみとし，特徴抽出は ST-GCN の代わりに GCN[10] を用いる．また，DINet の学習にはナビゲーションに適した視線を用いる．

4.3 異なる視線情報による比較

図 6(a) および図 6(b) に案内文生成結果の例を示す．図 6(a) に示すように，通常運転時の視線を用いた場合は対向車を対象としており，直進の指示に対して直感的に分かりづらい案内であることが確認できる．一方，ナビゲーションに適した視線を用いた場合前方車に対して追従する案内となっており，直感的に理解できる案内であることが確認できる．また，図 6(b) に示すように，通常運転時の視線を用いた場合は「human group」に対して「following」となっており，不適切な案内文であることが確認できる．一方，ナビゲーションに適した視線を用いた場合「following the ambulance」となっており，該当シーンでは救急車は直進しているため，不適切なナビゲーションであることが確認できる．しかし，ナビゲーションに適した視線を用いたことで，視覚的に発見しやすいと考えられる救急車を対象としている考えられる．したがって，ナビゲーションに適した視線は有効な手段であると言える．

4.4 時系列情報の有無による比較

図 6(c) および図 6(d) に生成結果例を示す．図 6(c) に示すように，時系列情報ありの場合は「following kawasaki ninja」と移動をともなう案内文が生成されている．したがって，動的なオブジェクトに着目したわかりやすいナビゲーションの生成できていることが確認できる．一方で，時系列情報なしの場合,「kawasaki ninja is located」と移動をともなわない案内文が生成されている．時系列情報がない場合でも，動的なオブジェクトに着目はできているが,「following」が「located」となっており，動きの情報が欠落していることが確認できる．また，図 6(d) の場合では，時系列情報ありの場合は進行先に対する行動が「following」であり，該当シーンにおいて不適切なナビゲーションになっていると確認できる．一方で，時系列情報なしの場合では，目標となる「human」の単語が繰り返しになっていることが確認できる．したがって，時系列情報の有無に関わらず図 6(d) の場合では案内文の生成が失敗していると言える．しかし，時系列情報ありの場合では,「human group」を生成できている．これは，時系列情報を導入したことにより各時刻における「human group」が紐づけられ，認識できたと考えられる．これらの結果から，時系列情報を導入することは HℓG において有効であることが確認できた．

自然言語生成の指標における定量的評価を表 1 に示す．表 1 より，時系列情報を導入することによって各評価指標での精度が向上しており，目的である Human-like Guidance に重要となる自然な案内文が生成されている．したがって，時系列情報の導入が有効な手段であることが確認できる．

表 1　時系列の有無による定量的評価比較

	時系列なし	時系列あり
Bleu_4	0.291	**0.325**
METEOR	0.199	**0.223**
ROUGE_L	0.450	**0.520**
CIDEr	2.649	**2.945**
SkipThoughtCS	0.458	**0.487**

5 おわりに

本研究では，HℓG のための，視線情報埋め込みシングラフと時系列情報を用いたナビゲーション生成手法を提案した．視線情報と時系列情報を用いることで，人間のような状況に合わせたナビゲーションの生成が実現できた．視線推定の評価実験では，ナビゲーションに適した視線が HℓG において有効であると確認できた．時系列情報の有無による評価実験では，時系列情報を用いることが HℓG において有効であることが確認できた．今後の課題としては，視線推定モデルの変更や HℓG に適切な評価指標の検討などが挙げられる．

参考文献

[1] Miki Tsutom, Inoue Takuya, Goto Yuta, Watanabe Masahiko, and Hayashi Hatsuo. Human-like local navigation system inspired by a hippocampal memory mechanism. *Brain-Inspired Information Technology*, 266:29–32, 2009.

[2] Sheng Yang, Lin, Guosheng, Jiang, Qiuping, Lin, and Weisi. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia*, 22(8):2163–2176, 2019.

[3] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 636–644, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 1–9, 2015.

[6] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, (1):61–80, 2009.

[7] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *In International Conference on Learning Representations (ICLR)*, 2016.

[8] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. *In International Conference on Learning Representations (ICLR)*, 2014.

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *In Advances in Neural Information Processing Systems (NIPS)*, 29:3844–3852, 2016.

[10] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *The International Conference on Learning Representations (ICLR)*, 2017.

[11] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Association for the Advancement of Artificial Intelligence (AAAI)*, (912):7444–7452, 2018.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *In Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.