Vol.2022-CVIM-230 No.4
2022/5/12

インスタンスセグメンテーションを用いたシャープなアテンションマップ生成による動作認識

仁田　智也[1,a)]　平川　翼[2]　藤吉　弘亘[2]　玉木　徹[1,b)]

概要：本論文では動画認識における説明可能性の向上のために，インスタンスセグメンテーションを利用して Attention Branch Network (ABN) が生成するアテンションマップをシャープなものにする Object-ABN を提案する．従来手法とは異なり，提案手法はコストの高い人間による修正や追加アノテーションを必要としない．UCF101 を用いた実験によって，提案手法は元の ABN の性能を保ちつつシャープなアテンションマップを生成できることを示す．

Action Recognition by Generating Sharp Attention Maps with Instance Segmentation

1. はじめに

画像認識や動画認識についての多くの研究が行われており，様々な用途に使われている．例えば自動運転においては画像認識は非常に重要であり，車載カメラに写る歩行者や他の車両などの物体の検出 [3], [7] や，道路の白線検出 [14]，車道と歩道を分離するセグメンテーション [8] などに利用されている．また工場の製品ラインにおける不良品の検知 [10] にも使われている．これまで人が目視で行なっていた検査を，画像認識を用いることでどの製品のどの部分に異常があるのかを高速で検出することが可能になり，コスト削減や生産速度向上にも貢献している．これらの例のように，単に画像が何であるのかを識別するだけでなく，識別対象が画像中の「どこに」あるのかを検出することが求められている．そのようなタスクが物体検出やセグメンテーションであるが，画像を識別するタスクにおいても，「どこに」注目してそのような識別結果が得られたのかを説明する必要性が近年求められている．深層学習モデルは通常ブラックボックスであり，なぜその識別結果が得られたのかが分からないことも多く，この状況を改善しようとする説明可能 AI の研究 [22] が盛んに行われている．

画像認識の結果を可視化するための一つの手法が，近年ではモデルの精度を向上させる手法として注目されているアテンション機構 [20] である．アテンションとは画像や画像中の注目するべき場所の重要度（重み）を定める機構であり，画像として表示される場合にはアテンションマップと呼ばれる．画像中のある場所へのアテンションの重みが大きいほど，認識結果へのその場所の重要度が大きいと解釈される．このような重要度の可視化の試みは CAM (Class Activation Map) [22] や Grad-CAM [15]，Score-CAM [17] など多数研究されている．これらの手法の主な目的は識別結果のための重要度の可視化であるが，この重要度自体を識別に利用するアテンション機構を用いた手法が近年多く提案されている．他に Attention Branch Network (ABN) [6]（図 1(a) 参照）は，識別のための perception ブランチに加えてアテンションを計算するアテンションブランチを利用して，アテンションの可視化と識別性能の向上を同時に行っている．

一般にアテンション機構は性能向上を目的としているため，可視化されたアテンションの重みが大きい領域と，人が見て重要だと思う領域は異なっていることもある．このようなギャップが大きい場合には，アテンションの可視化によって認識結果を説明しようとしても人の解釈と大きく異なっているため，説明可能性が低下してしまう．また人が重要だと思う部分の重みが小さいアテンションが獲得さ

[1] 名古屋工業大学
Nitech, Gokisocho, Showa, Nagoya 466–0061, Japan
[2] 中部大学
1200 Matsumotocho, Kasugai, Aichi 487-0027, Japan
[a)] t.nitta.635@nitech.jp
[b)] tamaki.toru@nitech.ac.jp

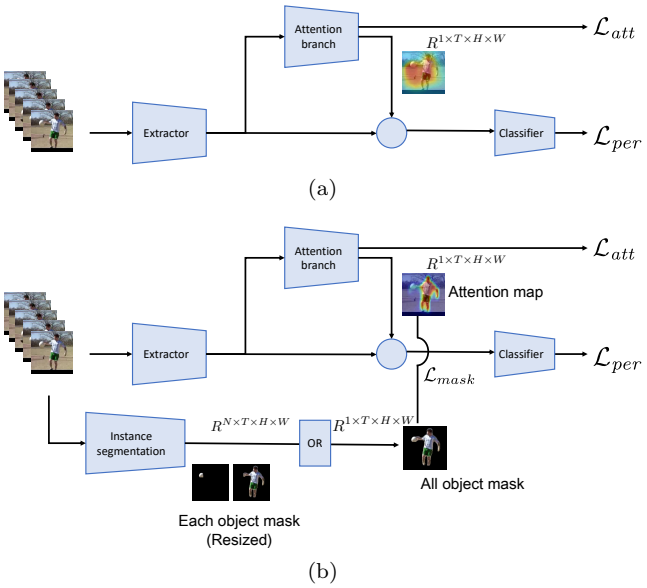

図 1 (a) ABN. (b) 提案手法である物体検出を用いた ABN.

れてしまい，汎化性能の低下にもつながることが懸念される．そこで，そのようなギャップのあるアテンションマップを修正する研究がいくつか存在する．三津原ら [23] は，モデルが生成したアテンションマップを人間が修正したアテンションマップと同様のマップを出力するよう学習することで，説明性の向上およびアテンション機構の性能を向上させている．しかし，人間によるアテンションマップの修正は画素単位のアノテーション作業であり，優れたユーザーインターフェースを利用したとしても，これを 1 枚ずつ行う人間が行うコストは非常に大きい．一方 Li ら [11] は，生成されたアテンションマップのどこに注目して欲しいのかを指定する追加のラベルを付与して，目的に沿ったアテンションマップを生成する方法を提案した．この追加ラベルには形状のセグメンテーションラベルが与えられているが，データセットが大規模になればやはりコストが大きくなる．

そこで本研究では，アテンションの修正を自動で行う手法である Object-ABN を提案する．これは ABN に基づいており，低コストで効果的にアテンションを修正することを可能にする．アテンションの修正を人手で行う従来手法 [23] とは異なり，提案手法ではセマンティックセグメンテーションの結果を利用する．人が見て重要な部分には，対象となる物体や人物が存在するはずである，ということを仮定し，画像中のその部分のアテンションがセグメンテーションで抽出された領域に近くなるようなマスク損失を設定する（図 1(b)）．

本研究の貢献は以下の通りである．

- インスタンスセグメンテーションを利用してアテンションマップを自動的に修正する Object-ABN を提案する．これはアテンション機構と人の解釈との重要度のギャップを埋めるものであり，手動で修正する従来手法と比べてコストが低い．
- 提案手法は，既存の様々なデータセットに対して適用できる．セマンティックセグメンテーションの結果を利用するため，タスクに依存せず，幅広い応用が期待できる．
- 複数のデータセットを用いた比較実験において，提案手法のようにアテンションへの損失を用いた場合でも性能が低下しないことを示す．またアテンションマップを可視化した場合の解釈性も向上することを示す．

2. 関連研究

2.1 アテンション機構

画像認識にアテンション機構を用いたモデルで有名な手法として Vision Transformer [2] がある．それ以前にも CNN モデルにアテンションを組み込んだ手法は多数提案されており [5], [9]，これらの手法ではアテンションはモデルの性能を向上させることを目的として使われている．

また，アテンションマップによる視覚的説明と分類精度の向上を目的とした手法として ABN [6] や ST-ABN [12] がある．この手法は中間特徴量から生成されるアテンションマップをアテンション機構に活用することで，説明性と認識性能の向上のどちらも満たすように設計されている．

動作認識へアテンション機構を組み込む研究はいくつか行われている．有名な手法として Non-Local Neural Network [18] がある．3 次元畳み込みニューラルネットワークをベースにした手法であり，フレーム内の空間的アテンションだけでなく，長期の時間方向のアテンションも利用する Non-Local block を提案した．

最近では Vision Transformer [2] を動画認識に適用した手法が複数提案されている [1], [13], [21].

2.2 アテンションの修正

生成されたアテンションマップはモデルの識別結果の根拠を視覚的に説明するものではあるが，実際には人間が注目する領域とアテンションが強い領域とが一致しない場合が生じる．またアテンションがかかる領域が非常に広く曖昧であり，シーン中のどの部分を注目しているか明瞭ではない場合もある．そこで，生成されるアテンションマップを人間の知見をもとに修正する手法が提案されている．三津原ら [23] は ABN を拡張することによって，人間がアテンションの修正をする Human-in-the-loop (HITL) の枠組みを提案した．また Guided Attention Inference Network (GAIN) [11] は，識別ラベルだけでなく，人間が着目した領域と同様のアテンションマップが出力されるよう，人間の注視領域を追加ラベルとして利用し，end-to-end で学習する方法を提案した．

モデルが生成するアテンションを HITL の枠組みで修正すれば，学習されたモデルでは人間が注目する場所と同じ

場所にアテンションがかかるようになるため，アテンションの可視化による説明性の向上が期待できる．しかし人間がアテンションを修正する方法はコストが大きく，大規模なデータセットに適用することは難しい．また GAIN のように追加ラベルとしてのアテンションマップを用意こともコストが大きい．さらに，これらの手法を動作認識タスクの動画像へ適用することを考えると，フレーム単位でアテンションマップのアノテーション作業が必要になり，どちらの手法も実用的ではない．

本研究の目的は動作認識の性能向上ではなく説明性の向上である．そのために可視化用のアテンションマップを生成する ABN [6], [12] をベースにして，動画像におけるアテンションの修正を自動化する．

3. 手法

本研究では動画認識のために 3 次元畳み込みを用いる Attetion Branch Network (ABN) [6], [12] を拡張する．従来の ABN と同様に損失はから得られる $\mathcal{L}_{att}$ と Perception Branch から得られる $\mathcal{L}_{per}$ の 2 種類の損失と，新しく提案するマスク損失 $\mathcal{L}_{mask}$ の 3 種類を用いて学習をする．本論文で新しく提案する損失は，モデルによって生成されたアテンションを，動画の各フレームに対してインスタンスセグメンテーションを適用して得られたマスクに近づけるものである．

3.1 ABN

ここでは動作認識に適用した場合の ABN の概要を説明する．

入力となる動画像を $x \in \mathbb{R}^{T_{\text{in}} \times 3 \times H_{\text{in}} \times W_{\text{in}}}$ とする．ここで T_{in} は動画像クリップのフレーム数，$H_{\text{in}}, W_{\text{in}}$ は動画像フレームの高さと幅である．ABN は特徴抽出器 E，アテンションブランチ A，識別器であるパーセプションブランチ P からなる．E から得られる中間特徴量を $h_1 = E(x) \in \mathbb{R}^{T \times C \times H \times W}$ とすると，アテンションブランチは h_1 を受け取り，アテンションマップ $M \in \mathbb{R}^{T \times 1 \times H \times W}$ と多クラス識別の予測ベクトル $y_m \in [0,1]^L$ を生成する．ここで L はクラス数である．交差エントロピー損失を L_{CE} として，アテンションブランチの出力と真値 y との損失を

$$\mathcal{L}_{att} = \mathcal{L}_{CE}(y_m, y) \tag{1}$$

とする．

アテンションマップ M は中間特徴量 h_1 に対して

$$h_2[:,c,:,:] = h_1[:,c,:,:]M[:,0,:,:] \tag{2}$$

もしくは

$$h_2[:,c,:,:] = h_1[:,c,:,:](1 + M[:,0,:,:]) \tag{3}$$

のように適用される．パーセプションブランチ P は $h_2 \in \mathbb{R}^{T \times C \times H \times W}$ を受け取り，多クラス識別の予測ベクトル $y_p = P(h_2) \in [0,1]^L$ を生成する．この識別器の損失を

$$\mathcal{L}_{per} = \mathcal{L}_{CE}(y_p, y) \tag{4}$$

とすると，最終的な損失は次式で表される．

$$\mathcal{L} = \mathcal{L}_{per} + \lambda \mathcal{L}_{att} \tag{5}$$

ここで λ は重みである．

3.2 Object-ABN

前述したとおり，アテンションブランチが生成するアテンションマップ M は，人の解釈とは異なった場所に大きな重みを持っていたり，広く曖昧に広がっていたりする．本研究では動作認識タスクにおいて，動画中に写る物体や人物領域が動作認識カテゴリを識別するために重要であると仮定し，アテンションマップの形状を物体領域や人物領域に近づける．そのために，事前学習済みのインスタンスセグメンテーションモデル S（具体的には Detectron2 [19]）を利用した自己教師あり学習による Object-ABN を提案する．このモデルに入力動画像 x の各フレームを入力し，インスタンス毎のマスク $M_s \in \{0,1\}^{T \times N(t) \times H \times W}$ を出力する．ここで $N(t)$ は時刻 t において検出されたインスタンスの個数である．なおこの $N(t)$ は可変であり，フレームの時刻 t 毎に異なる．

次にこのマスクを，次式に示す論理和を用いて 1 チャンネルのマスク $M'_s \in \{0,1\}^{T \times 1 \times H \times W}$ へ集約する．

$$M'_s[t,0,:,:] = \bigcup_{n=1}^{N(t)} M_s[t,n,:,:], t = 1, \ldots, T \tag{6}$$

これが望ましいアテンションであるとして，アテンションブランチが出力するマップ M との平均二乗誤差（MSE）をマスク損失

$$\mathcal{L}_{mask} = \mathcal{L}_{\text{MSE}}(M, M'_s) \tag{7}$$

として設定する．

最終的な損失は

$$\mathcal{L} = \mathcal{L}_{per} + \lambda \mathcal{L}_{att} + \lambda_{mask} \mathcal{L}_{mask} \tag{8}$$

であり，ここで λ_{mask} は重みである．

4. 実験

4.1 実験設定

データセット：実験に用いた UCF101 [16] は約 9500 動画の訓練セットと約 3500 動画の検証セットからなる人物動作 101 クラスの動作認識データセットである．各動画は Youtube から収集され，長さは短いもので 1 秒，長いもの

で 30 秒程度であるが，多くの動画の長さは 3 秒から 10 秒程度で，平均は 7.21 秒である．訓練と検証のスプリットが 3 種類あり，本研究では第 1 スプリットの性能を用いる．訓練用・検証用の動画数は (9537, 3783) である．

学習：訓練セットの 1 つの動画から連続する 64 フレームに対して 4 枚ごとに 1 フレーム，合計 16 フレームのクリップをサンプリングする．空間方向の短辺サイズを 256 画素から 320 画素の範囲でランダムに決定し，アスペクト比を保ったままリサイズをした後，224×224 画素をランダムに切り取り，一定の確率で水平反転を行う．学習に用いたオプティマイザは Adam，学習率は 10^{-4} に設定し，学習エポック数は 50 とした．

検証：検証セットの動画に対して学習時と同様に 1 クリップをサンプリングし，短辺サイズを 256 画素としてアスペクト比を保ったままリサイズをした後，中央部分 224×224 画素部分を切り取る．

アテンションマップの定量的評価：モデルが生成するアテンションマップの定量的評価のために，本研究ではエントロピーを用いる．

アテンションマップの値は 0 から 1 であるため，アテンションが画像全体にぼやけていればその分布は広くなり，エントロピーは大きくなる．一方で物体と背景にアテンションがシャープに分かれていれば分布は二極化し，エントロピーは小さくなるはずである．そこで，エントロピーが小さいほどシャープなアテンションマップであると言うことができる．ただしアテンションマップの値がある一定範囲内に収まるような，全体的にフラットなアテンションマップの場合にもエントロピーが低くなってしまう．そこで，ヒストグラムの最小値と最大値を 0 から 1 に正規化してからエントロピーを計算する．

本研究では区間の個数を $N = 10$ として，アテンションマップのヒストグラムを生成する．このヒストグラムの各区間 i の頻度 $hist[i]$ を正規化した離散確率 p_i からエントロピーを次式で計算する．

$$p_i = \frac{hist[i]}{\sum_{j=1}^{N} hist[j]} \tag{9}$$

$$\text{entropy} = \sum_{i=1}^{N} -p_i \log_2 p_i \tag{10}$$

ある動画の各フレームのアテンションマップに対してこのエントロピー計算を行い，その時間平均をその動画のアテンションマップのエントロピーとする．

アテンションマップのエントロピーは最大値は $i = 1, \ldots, 10$ について $p_i = 1/10$ の場合であり，このとき

$$\text{entropy} = \sum_{i=1}^{N} -\frac{1}{N} \log_2 \frac{1}{N} = \log_2 N \tag{11}$$

であり，今回の場合は $N = 10$ であるため $\log_2 10 \approx 3.332$ が最大値である．

表 1　UCF101 の検証セットに対する性能評価.

$\mathcal{L}_{per/attn}$	$\mathcal{L}_{mask}$	top-1	top-5	entropy	entropy object
✓		93.96	99.15	3.064	
✓	✓	93.62	99.26		2.026

モデル：提案手法は，X3D-M [4] をバックボーンとして ABN に拡張した．なお全ての実験において，X3D-M を Kinetics400 で事前学習した重みを初期値とした．

X3D-M を ABN に拡張する際に，X3D を前半部分と後半部分で分割し，前半部分である特徴抽出器の後に新しくとなる畳み込み層を追加し，アテンションを生成し適用した後，X3D の後半部分であるパーセプションブランチに特徴量が渡される．このモデルはサイズ $T \times H \times W$ の動画クリップを入力とし，サイズ $T \times H' \times W'$ のアテンションマップを生成する．ここで H', W' は特徴量の空間サイズである．以下の実験では $T \times H \times W = 16 \times 224 \times 224$, $T \times H' \times W' = 16 \times 14 \times 14$ とした．

パラメータ：実験に用いたパラメータは以下の通りである．$\lambda = 1, \lambda_{mask} = 10$.

比較：実験で比較する組み合わせを表 1 に示す．インスタンスセグメンテーションのマスク損失を加えたモデルと通常の ABN の比較を行う．

4.2　実験結果

マスク損失の効果を確かめるために，元の ABN に対してマスク損失を追加した Object-ABN と元の ABN を比較する．この結果は表 1 の上 2 行に対応する．この表から分かるとおり，マスク損失の有無で性能には大きな差が見られなかった．しかし生成されるアテンションマップは図 2(b)(c) に示すように全く異なっている．マスク損失を用いた場合（図 2(c)）には物体に対してシャープなアテンションマップが生成されているのに対し，マスク損失を用いていない場合（図 2(b)）は物体と背景の区別なくまだら模様のアテンションマップが得られている．またエントロピーも 1 以上低下しており，定量的にもアテンションマップがシャープになっていることが分かる．この結果から，マスク損失を用いることで性能の低下なしにシャープなアテンションマップを得ることができる事がわかる．

5.　おわりに

本論文ではインスタンスセグメンテーションによって ABN を拡張した Object-ABN を提案し，ABN よりもシャープなアテンションマップの生成を可能にし，深層学習による動画認識の判断基準をより明確にすることが可能となった．

ABN にマスク損失を加えた Object-ABN は，ABN と同等の性能を持ちながらシャープなアテンションマップの生

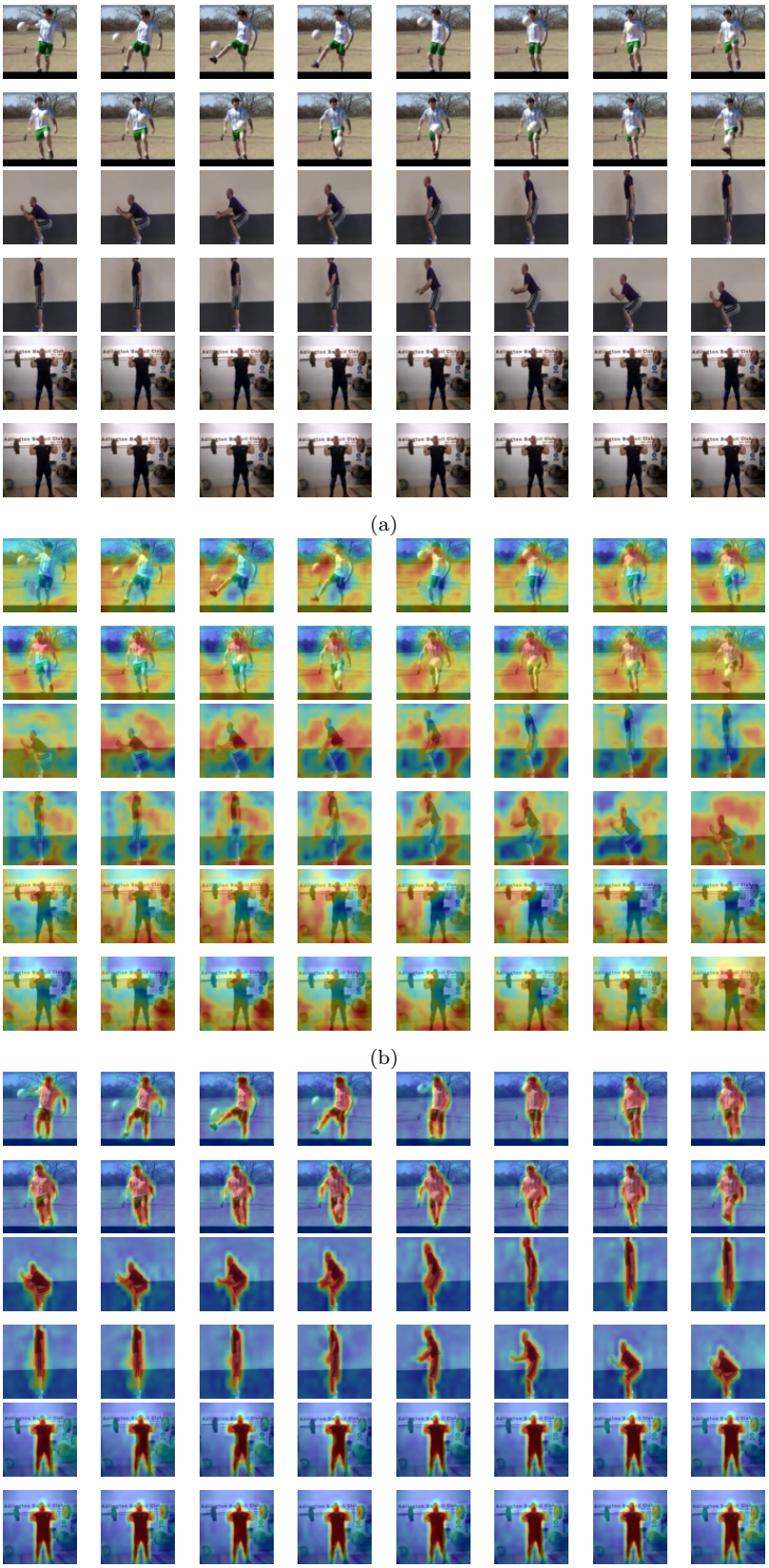

図 2 UCF101 の検証セットに対するマスク損失の効果．(a) 入力動画．(b) ABN で得られたアテンションマップ．(c) ABN にマスク損失を追加して得られたアテンションマップ．

成を行うことが可能であると分かった.

謝辞

本研究の一部は，JSPS 科研費 JP22K12090 の助成を受けた.

参考文献

[1] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. and Schmid, C.: ViViT: A Video Vision Transformer, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6836–6846 (2021).

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=YicbFdNTTy⟩ (2021).

[3] Fan, Q., Brown, L. and Smith, J.: A Closer Look at Faster R-CNN for Vehicle Detection, *2016 IEEE Intelligent Vehicles Symposium (IV)*, IEEE Press, p. 124–129 (online), DOI: 10.1109/IVS.2016.7535375 (2016).

[4] Feichtenhofer, C.: X3D: Expanding Architectures for Efficient Video Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

[5] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H.: Dual Attention Network for Scene Segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[6] Fukui, H., Hirakawa, T., Yamashita, T. and Fujiyoshi, H.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[7] Girshick, R., Donahue, J., Darrell, T. and Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[8] Hou, Y., Ma, Z., Liu, C., Hui, T.-W. and Loy, C. C.: Inter-Region Affinity Distillation for Road Marking Segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

[9] Hu, J., Shen, L. and Sun, G.: Squeeze-and-Excitation Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[10] Kim, T.-H., Kim, H.-R. and Cho, Y.-J.: Product Inspection Methodology via Deep Learning: An Overview, *Sensors*, Vol. 21, No. 15 (online), DOI: 10.3390/s21155039 (2021).

[11] Li, K., Wu, Z., Peng, K.-C., Ernst, J. and Fu, Y.: Tell Me Where to Look: Guided Attention Inference Network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[12] Mitsuhara, M., Hirakawa, T., Yamashita, T. and Fujiyoshi, H.: ST-ABN: Visual Explanation Taking into Account Spatio-temporal Information for Video Recognition, *CoRR*, Vol. abs/2110.15574 (online), available from ⟨https://arxiv.org/abs/2110.15574⟩ (2021).

[13] Neimark, D., Bar, O., Zohar, M. and Asselmann, D.: Video Transformer Network, *CoRR*, Vol. abs/2102.00719 (online), available from ⟨https://arxiv.org/abs/2102.00719⟩ (2021).

[14] Philion, J.: FastDraw: Addressing the Long Tail of Lane Detection by Adapting a Sequential Prediction Network, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[15] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

[16] Soomro, K., Zamir, A. R. and Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *CoRR*, Vol. abs/1212.0402 (online), available from ⟨http://arxiv.org/abs/1212.0402⟩ (2012).

[17] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P. and Hu, X.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2020).

[18] Wang, X., Girshick, R., Gupta, A. and He, K.: Non-Local Neural Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[19] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. and Girshick, R.: Detectron2, `https://github.com/facebookresearch/detectron2` (2019).

[20] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, *Proceedings of the 32nd International Conference on Machine Learning* (Bach, F. and Blei, D., eds.), Proceedings of Machine Learning Research, Vol. 37, Lille, France, PMLR, pp. 2048–2057 (online), available from ⟨https://proceedings.mlr.press/v37/xuc15.html⟩ (2015).

[21] Zhang, H., Hao, Y. and Ngo, C.: Token Shift Transformer for Video Classification, *CoRR*, Vol. abs/2108.02432 (online), available from ⟨https://arxiv.org/abs/2108.02432⟩ (2021).

[22] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A.: Learning Deep Features for Discriminative Localization, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[23] 三津原将弘, 福井宏, 坂下祐輔, 緒方貴紀, 平川翼, 山下隆義, 藤吉弘亘：Attention map を介した Deep Neural Network への人の知見の組み込み，電子情報通信学会論文誌, Vol. J104-D, No. 11, pp. 796–807 (2021).