社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

複数解像度間の整合性を考慮した顕著性予測のモデル効率化

瀬尾　俊貴† 　平川　翼† 　山下　隆義† 　藤吉　弘亘†

† 中部大学　〒487–8501 愛知県春日井市松本町 1200
E-mail: †{seotoshiki,hirakawa}@mprg.cs.chubu.ac.jp, ††{takayoshi,fujiyoshi}@isc.chubu.ac.jp

あらまし　顕著性とは人間の注視の引きつけやすさを意味し，画像から視覚特徴を抽出・統合することにより，顕著性マップとして人間が注意を向けやすい領域を求めることが可能である．そのため，動画像における顕著性予測では，動画要約やセマンティックセグメンテーションを始めとして様々な手法に用いられている．一般的に，顕著性予測は前述した動画要約やセマンティックセグメンテーションなどメインタスクの補助情報として用いられることが多い関係上，精度以外にも学習・評価時間やメモリ使用量の考慮が重要である．そこで，本研究ではモデル効率化を目指した顕著性予測，及び画像解像度毎の整合性を考慮した学習手法を提案する．提案手法では，MobileNetV2 をベースとしたモデルの利用によりモデルの効率化を行い，さらに中間特徴量を洗練する Refinement layer，画像解像度毎の顕著性の整合性を考慮することで，モデルのパラメータ数の軽減，及び精度が向上することを確認した．
キーワード　顕著性予測，深層学習，畳み込みニューラルネットワーク

Model efficiency of Salicncy prediction using consistency of multiple resolutions

Toshiki SEO†, Tsubasa HIRAKAWA†, Takayoshi YAMASHITA†, and Hironobu FUJIYOSHI†

† Chubu University　1200 Matsumoto, Kasugai, Aichi, 487–851 Japan
E-mail: †{seotoshiki,hirakawa}@mprg.cs.chubu.ac.jp, ††{takayoshi,fujiyoshi}@isc.chubu.ac.jp

Abstract　Saliency means the ease of attracting human gaze, and it is possible to extract and integrate visual features from images to create a prominence map of areas that people are likely to pay attention to. Therefore, video summarization and semantic segmentation are used in various methods for predicting saliency in video images. In general, since saliency prediction is often used as auxiliary information for the main task such as video summarization and semantic segmentation, it is important to consider the learning/evaluation time and memory usage in addition to accuracy. In this paper, we propose a learning method that takes into account prominence prediction and consistency at each image resolution to improve model efficiency. The proposed method reduces the number of parameters and improves the accuracy of the model by using MobileNetV2-based model, refinement layer for intermediate features and consistency of saliency at each image resolution.
Key words　Saliency prediction，Deep learning，Convolutional neural network

1. はじめに

動画像から人間が興味を示す領域を予測する顕著性予測という分野は，人間の視覚的特性の理解のために長年研究がされている．Itti らが提案した輝度・色相・エッジ等のボトムアップ性注意を抽出して予測するモデル [1] から始まり，人間の記憶や知見を活用するトップダウン性注意を導入したモデル [2] を経て，ボトムアップ・トップダウン性注意を統合して学習を行う Deep Learning を活用した手法に至り，目覚ましい発展を遂げている．[3]〜 [8] 近年では，RGB 画像に加えて Depth 情報や，音声情報を利用したモデルといった複数の入力情報から学習を行うマルチモーダル学習 [9] [10] や、動画要約やセマンティックセグメンテーション、自動運転といったメインタスクの補助的な情報として顕著性を利用した手法も活発に提案されている [11]〜 [14].

ただし，これらを組込みソフトウェアとして実装するにあたって，それらのモデルを用いるのはあまり現実的ではない．なぜなら、学習済みモデルによるメモリ使用量の圧迫や推論速度がボトルネックになるからである．そこで，本研究ではモバイル端末向けに実装された MobileNetV2 [15] を顕著性予測の分野に適用することでパラメータ数を削減する．しかしながら，単純にパラメータ数を削減すると精度が低下するた

め，Refinement layer の利用による特徴量の洗練、及び学習において各解像度毎の Saliency の整合性を利用した損失関数を定義するモデルを提案する．

2. 関 連 研 究

顕著性とは，人間が興味を示す領域を示したものであり，これを実現することができると様々な分野に応用することが可能となる．また，画像解像度を元の解像度から半分ずつ減少させていった時に，顕著性マップはどのような変化が現れるのかを被験者を集めて視線を取得し，調査した論文が存在する [20]．一方，モバイル端末向けに演算量を削減しつつ精度を維持する構造を用いた手法として MobileNetV2 がある．本章では，本研究に関連するこれらについて説明する．

2.1 顕 著 性

顕著性推定 (Saliency estimation) に関する研究は，画像上で人間の興味・関心のある領域を視線ベースで予測する顕著性予測 (Saliency prediction) と，画像内で特徴的な物体領域をピクセル単位で検出する顕著性検出 (Saliency object detection) があるが，前者の顕著性マップに関する研究は，1998 年の Itti らによって実用的な手法が提案されて以来，様々な手法が提案されてきた．人間は視覚から得られる情報をもとに注目する領域を決定する際に，第一次視覚野によって色やエッジ，方向といった低次の特徴の処理が行われる．この知見に基づいて従来の手法は，色やエッジ，方向といった低次特徴（ボトムアップ性注意）を用いたボトムアップ型の処理を用いている．また，顕著性を決定する大きな要因として画像内に人間や顔が存在するとき人間や顔に注目する傾向があるなどといった高次特徴（トップダウン性注意）がある．そこで，従来の手法では顔領域に対して顕著性を高める手法が次に提案されている．しかしながら，注視領域を決定する要因は記憶や経験などから得られる様々な要因があり，同一物体でも異なるシーンにおいては顕著度が変わる可能性もあるため，これらすべての特徴量を表現することは非常に困難であった．そのため，顕著性予測においては Convolutional Neural Network (CNN) による手法が大きく成果を上げている．CNN では，浅い層では，エッジや色，方向といった低次特徴に反応し，層が深くなるにつれて，高次特徴に反応することが知られている．これによって，ボトムアップ・トップダウン性注意を統合して学習を行い，かつ多様な顕著性を近似することが可能となった．さらに近年では，RGB 画像に加えて Depth 情報や，音声情報を利用したモデルといった複数の入力情報から学習を行うマルチモーダル学習や，動画要約やセマンティックセグメンテーション，自動運転といったメインタスクの補助的な情報として顕著性を利用した手法も提案されている．

2.2 顕著性と解像度に関する研究

各解像度と顕著性マップにどのような関係性があるのかを，実際に被験者から視線を集めて調査した研究がある．[20] これによると，比較的解像度の高い時の顕著性マップは元の解像度の顕著性マップと比べて，非常に類似度の高い顕著性マップが得られ，逆に解像度が極端に低い顕著性マップになるほど類似度が低くなっている．ただし，この研究ではこの相関が CNN にどのような影響を与えるのかの検証は行っていない．そのため，事前調査として CNN モデルを利用して同じ顕著性マップの真値を異なる解像度の画像から予測した時に，どのような精度変化が現れるかを図 1 に示す．ここで，それぞれ 1/1 は元の解像度，1/2 は元の解像度から 1/2 の解像度へリサイズ，1/4 は元の解像度から 1/4 の解像度へリサイズ，1/8 は元の解像度から 1/8 の解像度へリサイズ，1/16 は元の解像度から 1/16 の解像度へリサイズ，1/32 は元の解像度から 1/32 の解像度へリサイズ，1/64 は元の解像度から 1/64 の解像度へリサイズした画像を読み込み学習・評価した結果である．

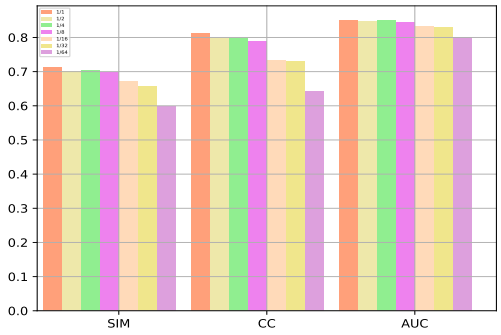

図 1　CNN モデルによる予測精度の事前調査

図 1 のように，解像度が 1/8 まではほぼ同等の精度を維持しているが，1/16 より下回る解像度になると精度は大幅に低下する．これは，画像としての情報が単調になりすぎるため，画像としての特徴が殆ど同一になってしまうためであると考えられる．この傾向は，文献 [20] でも同様に表れている．

2.3 MobileNetV2

モバイル端末向けの CNN モデルとして提案された手法として MobileNetV2 がある．MobileNetV2 の構造を図 2 に示す．

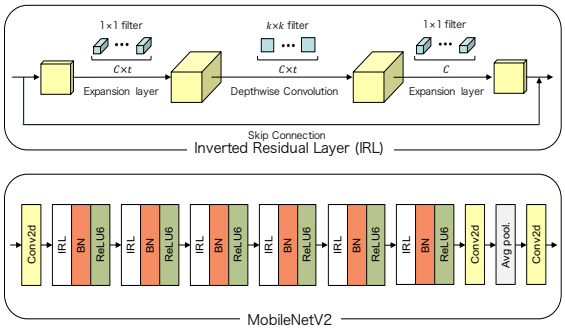

図 2　MobileNetV2 の構造

一般的な畳み込み処理では，出力特徴マップのチャンネル数が C なら，畳み込みフィルタの枚数が C 必要になる．各フィルタはチャンネル方向にも軸が存在するため，フィルタの枚数が増えると計算コストやパラメータ数が膨大になる．そのため，フィルタサイズを 1 にして畳み込み処理を行う Pointwise Convolution を行ったのち，フィルタの枚数を 1 枚にして Convolution を行う Depthwise Convolution を行い，空間・チャンネル方向の畳み込みを分けて処理することで演算量を削

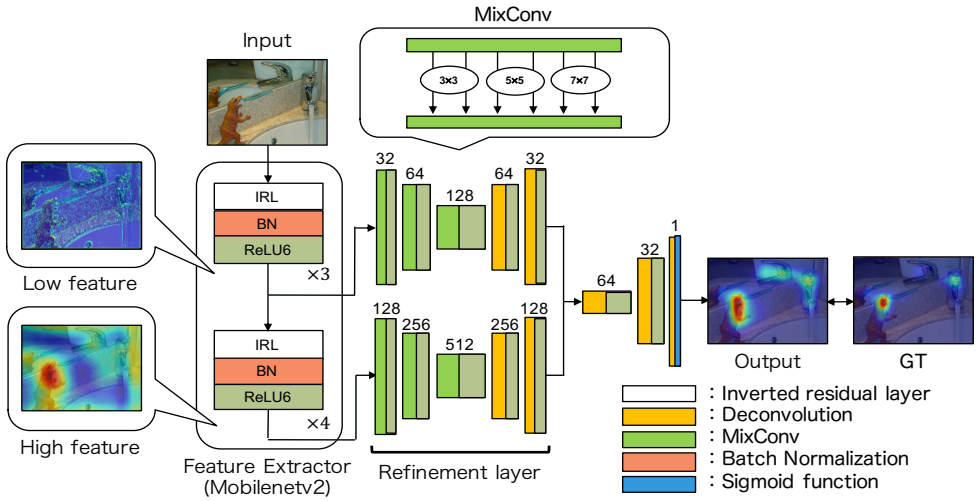

図 3　提案手法のネットワーク構造

減することが可能である．ただし，演算量を減らすことは表現力を犠牲にすることと等価であるため，MobileNetV2 では，入力チャンネルを控え目に設定し，Pointwise Convolution の際に中間特徴量である出力チャンネルを t 倍に写像する Expantion layer を用いて畳み込み処理を行い，Depthwise Convolution の後に入力チャンネルと同じチャンネル数へ戻すことで表現力を落とさずに精度を維持することが可能となった．また，勾配消失の問題も兼ねて Skip Connection [16] を導入している．そして，Expansion layer の拡張率 t は通常 6 にしている．

表 1　Inverted residual layer の各ハイパーパラメータ

Number of layer	t	c	n	s
1	1	16	1	1
2	6	24	2	2
3	6	32	3	2
4	6	64	4	2
5	6	96	2	2
6	6	160	3	2
7	6	320	1	1

3. 提 案 手 法

従来手法ではモデルのパラメータ数は膨大となり，推論速度が低下する傾向があるため，顕著性予測を利用したメインタスクを実行する場合ではメモリのリソースや処理速度の観点からあまり現実的ではない．そこで，本研究ではパラメータを削減するために MobileNetV2 や Mixed Depthwise Convolution [17] を利用したネットワーク構造の導入を行う．また，精度の維持も行うためにボトムアップ・トップダウン性注意の洗練を行う Refinement Layer 及び各解像度間の顕著性の整合性を考慮した損失関数を提案する．ここでは，提案手法の構築方法，損失関数の定義・学習の詳細について説明する．

3.1　ネットワーク構造

提案手法のネットワーク構造を図 3 に示す．各ブロックに記載してある数値は出力チャンネル数を表す．始めに，Inverted residual layer による計算コストと勾配消失を軽減した畳み込み処理を用いた MobileNetV2 を特徴抽出器として画像から特徴を得る．この時，MobileNetV2 は Inverted residual layer を合計 10 層配置した構成となっており，前半の 3 層と後半の 4 層から特徴を得る．また，Inverted residual layer には入力チャンネルの拡張率 t，出力チャンネル数 c，ブロックの繰り返し数 n，畳み込み時のストライド幅 s を各層毎にハイパーパラメータとして設定する．提案手法におけるハイパーパラメータを表 1 に示す．

前半の特徴では，ボトムアップ性注意 (周りとは違う色情報や識別に有用なエッジなどの単純な視覚情報) を，後半の特徴ではトップダウン性注意 (人間が成長する上で獲得した知見に基づいた視覚情報) が得られる．その後，得られた特徴を Refinement layer に入力する．Refinement layer は得られた特徴を洗練するモジュールとなっており，入出力の解像度を変更しない Encoder-Decoder 構造から構成されている．さらに，Encoder-Decoder 構造には Mixed depthwise Convolution (MixConv) を用いることで，演算量を抑えつつも様々な受容野を利用することでロバストな特徴の抽出が可能である．一般的な CNN は，単純にフィルタサイズを大きくほど精度が向上するが，その分パラメータ数も増加する．具体的に，CNN におけるパラメータ数は，画像サイズを $H \times W$ 入力チャンネル数を N，出力チャンネル数を M，フィルタサイズを k とすると，$H \times W \times N \times M \times k^2$ と表せる．そのため，フィルタサイズが大きくなるほど演算量も増加する．そこで，MixConv ではパラメータ数をなるべく抑えつつ精度を向上させるために，チャンネルにグループを設定し，グループ毎に異なるフィルタサイズの畳み込み処理を行う．グループ数を 3，カーネルサイズを k_1，k_2，k_3 とした時，演算量は $H \times W \times N \times M \times (k_1^2 + k_2^2 + k_3^3)$ となり，計算量オーダーをべき乗に増やすことなく様々な受容野を用いた畳み込み処理を行うことが可能である．また，演算量の観点から見ると，Dilated Convolution を用いる選択肢もあったが，Dilated Convolution は大きいサイズのフィルタより精度が低下することが報告されているため，今回は MixConv を採用した．

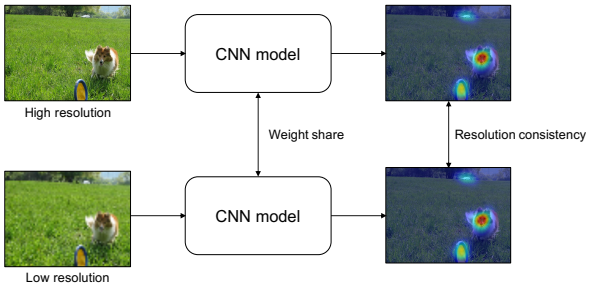

図 4　各解像度間の整合性を考慮した Consistency loss の設計

提案手法では，グループ数を 3 とし，フィルタサイズをそれぞれ 3，5，7 としている．最後に，Refinement layer から得られた 2 つの特徴を結合し，Deconvolution 層を用いて最終的な推論を行う．

3.2　損 失 関 数

従来研究 [20] によると，元の解像度とそれより低い解像度の顕著性を比較した時，ある程度までは解像度毎の相関が非常に高いことが知られている．また，事前調査として元の解像度から取得された顕著性マップを低解像度化した画像にも CNN モデルに与えて学習を行ったところ，1/8 までの解像度ではほぼ同等の精度を得られ，1/16 以降では大幅に精度が低下した．各解像度間の精度が前後する理由として，低解像度ほどボトムアップ性注意が，高解像度ほどトップダウン性注意が特徴的にわかりやすく含まれているからであると考えられる．従来研究で実際に各解像度に対して顕著性マップを取得した際にもある一定の解像度の顕著性マップは元の顕著性マップと非常に近くなり，一定以下の低解像度であると一致しなくなる．従って，各解像度毎の顕著性を一致させるために，図 4 に示すようにモデルに元の解像度と低解像度の画像を入力した時，出力される顕著性同士の差を最小化するように誤差関数を定義する．これにより，画像のシーンの認識に重要な大まかな色情報と元の解像度に存在する特定物体に対する特徴の両方を学習することが可能であると考えられる．そこで，本研究では Binaliy Closs Entropy (BCE) を利用した損失関数 L_{res} を式 (1) のように定義する．ここで，i は各ピクセル，P_h は解像度を 1/2 にした時の推論結果，P_q は解像度を 1/4，1/8 にした時の推論結果である．異なる解像度のコンテキスト情報を含むネットワークを学習後，推論の際は各解像度毎の画像を入力し，各推論結果の平均を最終出力とする．また，学習促進のため，あらかじめ元の解像度と顕著性マップの真値の BCE から学習したモデルを用いて最適化を行う．

$$L_{res} = -\sum_i \{S_i^{P_h} \log S_i^{P_q} + (1 - S_i^{P_h}) \log(1 - S_i^{P_q})\} \qquad (1)$$

4.　評 価 実 験

提案手法の有効性を確認するために，評価実験を行う．評価実験では，SALICON 及び CAT2000 データセットを用いて，顕著性マップの性能を評価する．

4.1　実 験 概 要

学習時におけるバッチサイズを 10 とし，エポック数を 100，最適化手法には Momentum SGD を用いる．また，初期学習率を 0.01 とし，30，50 エポックで 0.1 を乗算する．そして，L_{res} を用いた学習の際，低解像度化した画像はバイリニア補間により元の解像度へとアップサンプリングして入力する．評価データには，SALICON データセット [18] を用いる．SALICON データセットは，様々なシーンから構成される MS-COCO データセットから 20,000 枚を抜粋し，顕著性を付与したものである．顕著性は，各画像に対し約 60 名の被験者を対象に，マウスによるトラッキングデータから算出され，Amazon Mechanical Turk を利用して自動収集されている．そして， CAT2000 [19] データセットは 20 個のシーンを用いたデータ合計 4,000 枚に対して約 20 名の被験者から，角膜反射法を利用して視線から顕著性が収集されている．画像サイズは，SALICON データセットが 640 × 480，CAT2000 が 1920 × 1080 である．

SALICON データセットは学習用に 10,000 枚，評価用に 5,000 枚を用いる．また，CAT2000 データセットの顕著性マップの真値は一般公開されていないため，学習データ 2,000 枚を分割し，学習に 1,600 枚，評価用に 400 枚を用いる．そして，学習の高速化・汎化性能の観点から学習データに対して [0,1] への正規化，及び左右反転の Data Augmentation 処理を行っている．また，評価に利用する従来手法・提案手法を以下のように定義する．

SalNet : SalNet は, ImageNet [21] で事前学習された VGG16 [22] を利用して End-to-End で顕著性予測を行う手法である．

EML – Net : EML-Net は，複数の事前学習済みモデルを活用した学習を行うことで，事前にエンコードされた識別に有効な特徴を効果的に利用する手法である．

Vanilla : Vanilla は，Mobilenetv2 のデコード部を全結合層から Deconvoliution 層を 3 層に変更することで，顕著性予測用にネットワークを変更したモデルである．

Vanilla + R : このモデルは, Refinement layer 及び Mixed Depth-wise Convolution を活用することによりボトムアップ・トップダウン性注意の洗練を行なった後，それらの特徴を統合して Deconvolution 層へ入力するモデルである．

Vanilla + R + L_{res} : Vanilla+R のモデルに対して，各解像度の整合性を考慮した損失関数 L_{res} を導入したモデルである．

Vanilla+R+L_{res} は，元の解像度の顕著性マップを用いて損失を取る関係上，Vanilla のモデルから Fine-tuning を行い学習をする．

4.2　顕著性マップの評価

顕著性予測における定量的評価指標として，双方の各ピクセルの顕著度の最小値を合計した指標である SIM(Similarity)，双方の顕著性マップの相関係数を表す指標である CC(Correlation coefficient)，ROC 曲線から計算した指標である AUC(Area Under Curve)，双方の顕著性マップの直流成分を無視した指標である NSS(Normalized Scanpath Saliency)，双方の顕著性マップの分布の近さを表現した指標である KL(KL divergence) がある．これらの評価指標を用いて，SALICON データセットと CAT2000 データセットにおける各精度の比較結果を表 2，3 に示す．提案手法では，整合性を保つ損失関数の利用と Refinement Layer

により，従来手法と比較して精度が向上していることがわかる．また，MobileNetV2 と Mixed Depthwise Convolution を利用したモデルの利用により，パラメータ数を 87.7% 圧縮できていることが確認できる．そして，図 6 に各手法の定性的結果を示す．図 6 より，Vanilla では岩やキーボード，毛並みなどエッジ等の特徴が密集する画像において誤認識を誘発しやすかったが，Refinement layer と L_{res} の導入により不必要な特徴量を鮮明化することで正しい顕著性が推定できていることが確認できる．

表 2　SALICON データセットによる定量的評価結果

	Params	SIM ↑	CC ↑	AUC ↑	NSS ↑	KL ↓
SalNet	24.03M	0.702	0.793	0.847	1.621	0.382
EML-Net	47.08M	**0.765**	**0.878**	**0.864**	**1.987**	0.520
Vanilla	**4.72M**	0.686	0.779	0.844	1.577	0.393
Vanilla+R	5.75M	0.714	0.811	0.850	1.673	0.338
Vanilla+R+L_{res}	5.75M	0.725	0.830	0.860	1.704	**0.322**

表 3　CAT2000 データセットによる定量的評価結果

	Params	SIM ↑	CC ↑	AUC ↑	NSS ↑	KL ↓
SalNet	24.03M	0.752	0.841	0.859	1.752	0.315
EML-Net	47.08M	**0.780**	**0.886**	**0.866**	**2.050**	**0.298**
Vanilla	**4.72M**	0.746	0.833	0.858	1.731	0.321
Vanilla+R	5.75M	0.757	0.842	0.860	1.799	0.311
Vanilla+R+L_{res}	5.75M	0.770	0.850	0.863	1.870	0.304

4.3　モデル量子化による圧縮率の比較

提案手法では，MobileNetV2 や MixConv を利用することでパラメータ数を大幅に削減したが，モバイル端末や CPU による実行を想定すると改善の余地が残っていると考えられる．そこで，各パラメータのビット数を通常の 32 ビットから 16 ビット，8 ビットへと変更した際に，パラメータ数や精度がどのように変化するかを確認した結果を表 4 に示す．なお，データセットは SALICON，モデルは Vanilla+R+L_{res} を使用する．パラメータ数を比較すると，32 ビットでは 5.75M であったのが，16 ビットでは 3.10M，8 ビットでは 1.78M とそれぞれ 46.0%，69.0% 削減できていることが確認できる．ただし，量子化のみであると精度が低下する傾向があり，トレードオフの関係となっているため，元の精度を限りなく保つためにモデルの層毎の最適な量子化ビット数の探索等が必要であると考えられる．

表 4　モデル量子化による圧縮率の比較：SALICON

	Params	SIM ↑	CC ↑	AUC ↑	NSS ↑	KL ↓
32 ビット	5.75M	**0.725**	**0.830**	**0.860**	**1.704**	**0.322**
16 ビット	3.10M	0.711	0.818	0.855	1.659	0.339
8 ビット	**1.78M**	0.692	0.793	0.845	1.583	0.366

4.4　MixConv の導入による傾向の調査

提案手法では Refinement layer に対して異なるサイズの受容野から特徴を効率的に抽出するために，MixConv を導入している．そのため，Refinement layer に MixConv を導入する場合と，フィルタサイズを 3 に固定して通常の畳み込み処理を行った時の精度，及びパラメータ数を比較した結果を表 5 に示す．パラメータ数を比較すると，MixConv 有の場合に比べ，MixConv 無はパラメータ数が 51.4% 増加している．また，様々なフィルタサイズの畳み込み処理により，様々な空間的特徴が得られていることから，ほぼ同等の精度となっていることが確認できる．

表 5　MixConv の有無による評価結果：CAT2000

	Params	SIM ↑	CC ↑	AUC ↑	NSS ↑	KL ↓
MixConv 有	**5.75M**	**0.770**	**0.850**	**0.863**	**1.870**	**0.304**
MixConv 無	8.71M	0.767	0.847	0.861	1.859	0.311

4.5　中間特徴量の可視化

Refinement layer による中間特徴量が入力前後でどのような変化が起こっているのかを可視化した例を図 5 に示す．なお，各特徴の全体的な発火度合を定性的に示すために，全体の特徴マップの平均を算出したマップを出力している．また，左上が前半の 3 層から出力された主にボトムアップ性注意が出力される Low feature，左下が後半の 4 層から出力された主にトップダウン性注意が出力される High feature，右上・右下が Low feature・High feature を Refinement layer に入力した時の出力である．図 5 をみると，Refinement layer に入力する前の Low feature は周りの画素とは違う色やエッジが抽出されているが，全域に渡り反応している領域が多い．High feature も人間や花壇など特徴的な部分に対して強く反応していることが確認できるが，全域に渡り反応している領域が多い．そして，Refinement layer の入力後では，エッジや色・特定物体に対する反応がかなり特定できていることが確認できる．

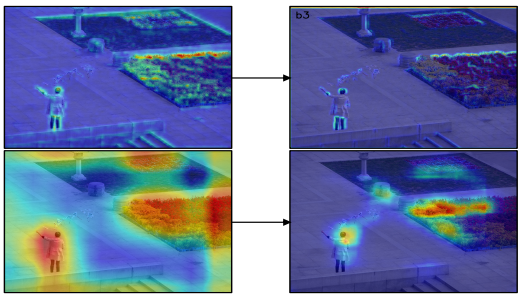

図 5　Refinement layer の前後による中間特徴量の可視化．左上：入力前の Low feature，右上：入力後の Low feature，左下：入力前の High feature，右下：入力後の High feature

5.　おわりに

本研究では，MobileNetV2 や Mixed Depthwise Convolution をベースとしたネットワークの設計によりパラメータ数を軽減した．また，各解像度間の顕著性の整合性を保つ損失関数の提案により，整合性を保つ損失関数を利用しない場合と比べ精度が向上することを確認した．今後は，更なるパラメータ

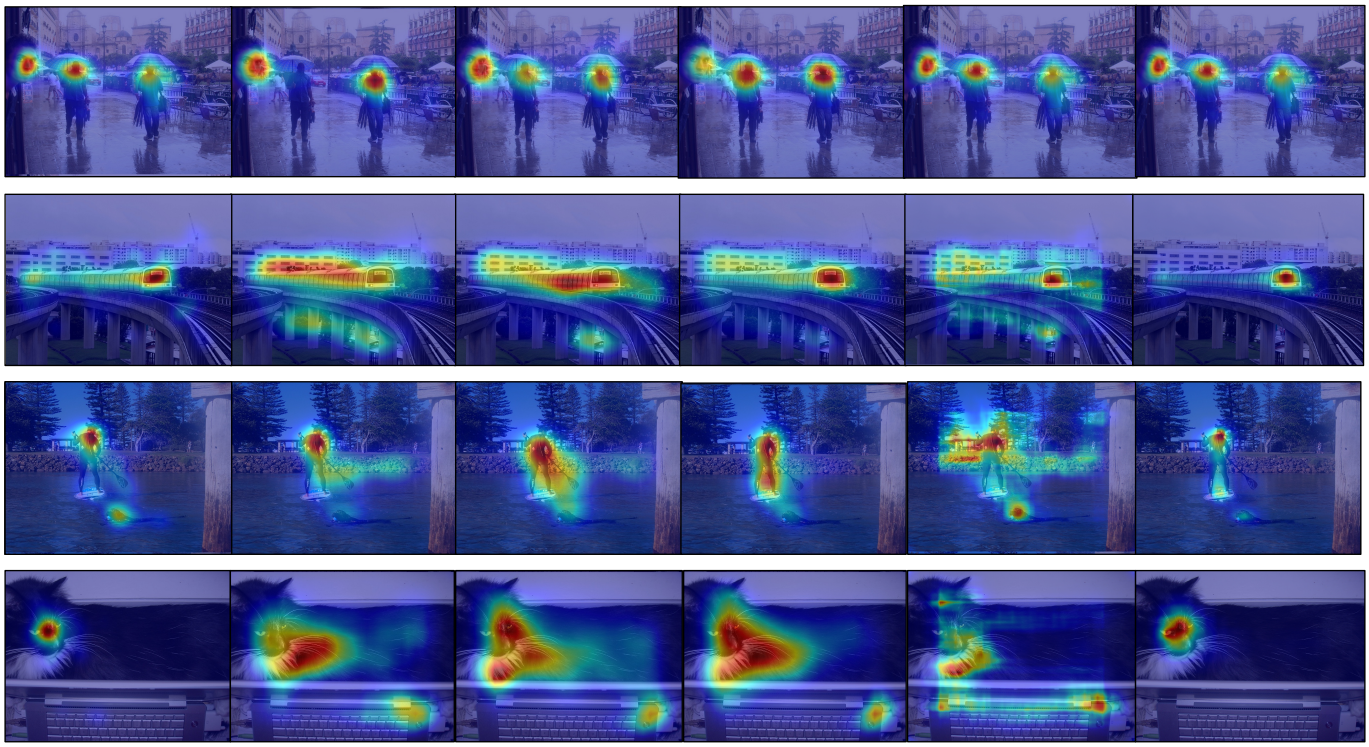

図 6　SALICON データセットを用いた定性的結果例

削減のために枝刈りや蒸留を導入した研究を行う.

文　　献

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.20, no. 11, pp.1254-1259, 1998.

[2] Gao, Dashan, and Nuno Vasconcelos. "Bottom-up saliency is a discriminant process." 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007.

[3] Pan, Junting, et al. "Shallow and deep convolutional networks for saliency prediction." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] M. Kummerer, L. Theis, and M. Bethge. "Deep gaze i: Boost- ¨ ing saliency prediction with feature maps trained on imagenet." In International Conference on Learning Representations, 2015.

[5] Li, Guanbin, and Yizhou Yu. "Visual saliency based on multiscale deep features." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[6] Huang, Xun, et al. "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[7] Jia, Sen, and Neil DB Bruce. "Eml-net: An expandable multi-layer network for saliency prediction." Image and Vision Computing (2020): 103887.

[8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. CoRR, abs/1611.09571, 2016.

[9] Tsiami, Antigoni, Petros Koutras, and Petros Maragos. "STAViS: Spatio-Temporal AudioVisual Saliency Network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[10] Liu, Nian, Ni Zhang, and Junwei Han. "Learning Selective Self-Mutual Attention for RGB-D Saliency Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[11] 片岡香織, 数藤恭子, and 森本正志. "室内構造推定と Saliency Map とを用いた看板検出技術." 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 111.222 (2011): 31-35.

[12] Wang, Wenguan, Jianbing Shen, and Fatih Porikli. "Saliency-aware geodesic video object segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Lee, Yong Jae, Joydeep Ghosh, and Kristen Grauman. "Discovering important people and objects for egocentric video summarization." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.

[14] Xia, Ye, et al. "Periphery-fovea multi-resolution driving model guided by human attention." The IEEE Winter Conference on Applications of Computer Vision. 2020.

[15] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[16] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Tan, Mingxing, and Quoc V. Le. "Mixconv: Mixed depthwise convolutional kernels." arXiv preprint arXiv:1907.09595 (2019).

[18] Jiang, Ming, et al. "Salicon: Saliency in context." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[19] Borji, Ali, and Laurent Itti. "Cat2000: A large scale fixation dataset for boosting saliency research." arXiv preprint arXiv:1505.03581 (2015).

[20] Judd, Tilke, Fredo Durand, and Antonio Torralba. "Fixations on low-resolution images." Journal of Vision 11.4 (2011): 14-14.

[21] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[22] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).