異なる解像度間の整合性を考慮した顕著性予測モデルに関する研究

TP19010 瀬尾俊貴　　指導教授：山下隆義

1. はじめに

顕著性予測は，人が興味・関心を持つと考えられる領域をヒートマップにより表現する．顕著性予測のアプローチとして，周囲とは異なる色彩やエッジ特徴を抽出して顕著領域を推定する手法がある．また，Convolutional Neural Network (CNN) の発展により，人の視線から得られた顕著性マップを学習・推論に用いる高精度な顕著性予測手法も提案されている．一方で，顕著性予測の応用先として自動運転システムを想定すると，予測精度だけではなくメモリ消費量や計算時間の短縮が重要となる．そこで，本研究ではメモリ消費量を効率化した顕著性予測モデルを提案する．さらに，パラメータの少ないモデルで高精度に推定するために，画像解像度毎の顕著性の一貫性を考慮した学習手法も提案する．

2. 顕著性予測

顕著性予測の代表的な手法として，RARE2012 [1] が挙げられる．RARE2012 では，主成分分析によりカラーチャンネル毎に白色化を行う．その後，得られた画像に対してマルチスケールのガボールフィルタで特徴量を抽出する．最後に，得られた特徴量の確率密度を計算し，自己情報量に基づいて顕著性予測を行う．ただし，RARE2012 では危険予知など人の事前知識にもとづく顕著性の違いを考慮できないという問題がある．また，人の視線情報を利用した CNN による高精度な顕著性推定手法がある．SalNet [2] では，畳み込み層 3 層，全結合層 2 層から構成されるネットワークを用いて顕著性マップを出力する．ネットワークの学習には人の視線データをもとにしたマップを教師として用いる，また，EML-Net [3] では，異なるデータセットで事前学習した複数ネットワークの視覚特徴を統合することで高精度な予測が可能である．しかし，EML-Net は，大規模なネットワークを利用して特徴抽出を行うため，モデルのパラメータ数が膨大となり，計算時間がかかる問題がある．

3. 提案手法

本研究では，メモリ消費量及び計算時間の短縮と予測精度の向上を両立する顕著性予測手法を提案する．まず，メモリ消費量と計算時間の短縮のために，MobileNetV2 と Mixed Depthwise Convolution を利用してネットワークを構築する．これにより，パラメータ数を削減できる．次に，予測精度を向上させるために，人の視覚的特性に基づいた解像度毎の視覚的特徴を効率よく学習する損失関数を提案する．

3.1 ネットワーク構造

提案手法で用いるネットワーク構造を図 1 に示す．はじめに，MobileNetV2 を特徴抽出器として画像から特徴を得る．この時，MobileNetV2 は IRL，Batch Normalization (BN)，ReLU6 を組み合わせた計 7 層の構成であり，前半の 3 層と後半の 4 層から特徴を得る．

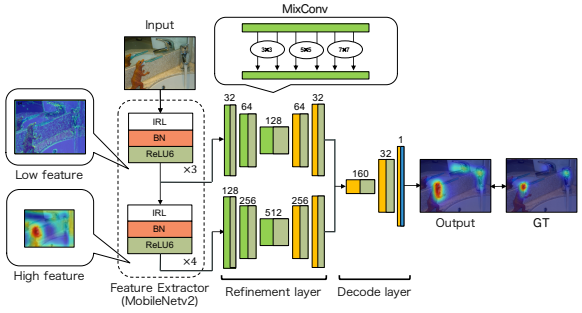

図 1：提案手法のネットワーク構造

前半の特徴は，浅い層から得られる特有の特徴である周囲とは異なる色情報や識別に有用なエッジなどの単純な視覚情報となる．また，後半の特徴は，深い層から得られる特有の特徴である物体などの大域的な視覚情報となる．次に，得られた特徴を Refinement layer に入力する．Refinement layer は得られた特徴を明瞭化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，Refinement layer には様々な受容野を 1 度の畳み込み層に集約を行った Mixed-depthwise Convolution (MixConv) を用いることで，演算量を抑えつつもロバストな特徴の抽出を行う．最後に，Refinement layer から得られた 2 つの特徴を統合し，Deconvolution 層を用いて顕著性マップの推論を行う．

3.2 損失関数

画像解像度と顕著性マップの関係性について，人の視線を収集し調査した研究がある [4]．これによると，解像度を 32 分の 1 までリサイズした画像の顕著性マップは元の解像度の顕著性マップと非常に相関が高いことが確認されている．この知見を活かし，元の解像度から予測される顕著性マップと低解像度化した画像から予測される顕著性マップの整合性を損失として求める．これを定式化すると式 (1) のように定義できる．

$$L_{res} = BCE(S^{GT}, S^{P_b}) + \sum_{i=1}^{N} BCE(S^{P_b}, S^{P_{2^i}}) \quad (1)$$

第一項は，バイナリクロスエントロピーを用いた真値と元の解像度から予測される顕著性マップの損失である．第二項は，元の解像度から予測される顕著性マップと低解像度化した画像から予測される顕著性マップからの損失である．ここで，BCE はバイナリクロスエントロピー，S^{GT} は顕著性マップの真値，S^{P_b} は元の解像度から予測された顕著性マップ，$S^{P_{2^i}}$ は解像度を 2^i 分の 1 にした時の推論結果，N はダウンサンプルする回数である．図 2 に示すように，低解像度の画像は元の解像度の画像を 2^i 分の 1 にダウンサンプリング後，バイリニア補間により元の解像度へアップサンプリングした画像とし，モデルに入力する．

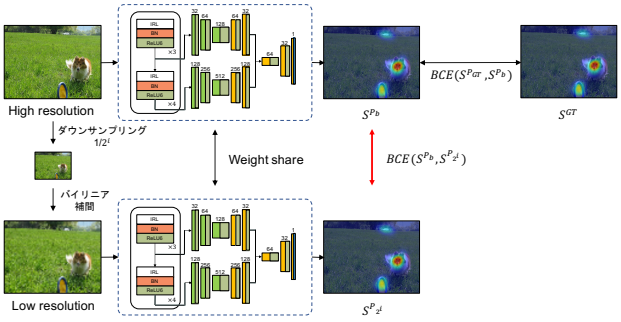

図 2：各解像度間の整合性を考慮した一貫性損失

4. 評価実験

提案手法の有効性を示すために，静止画における人の視線データから得られた顕著性を収集したデータセットを用いて，評価実験を行う．

4.1 実験概要

評価データには，SALICON 及び CAT2000 を用いる．SALICON は，様々な自然画像から構成される MS-COCO データセットから 20,000 枚を抜粋し，約 60 名の被験者から顕著性を獲得して付与したものである．また，CAT2000 は 120 名の被験者を対象に，漫画，アート，オブジェクト，低解像度画像，屋内，屋外，ランダムな画像，線画といった異なるタイプのシーンをカバーし，20 のカテゴリから構成されている．画像の入力サイズは 640 × 480 で

表 1：SALICON，及び CAT2000 データセットによる定量的評価結果

	SALICON						CAT2000					
	Params	SIM ↑	CC ↑	AUC ↑	NSS ↑	KL ↓	Params	SIM ↑	CC ↑	AUC ↑	NSS ↑	KL ↓
EML-Net[1]	47.08M	**0.765**	**0.878**	**0.864**	**1.987**	0.520	47.08M	**0.782**	**0.885**	**0.866**	**2.060**	**0.298**
Vanilla	**4.72M**	0.686	0.779	0.844	1.577	0.393	**4.72M**	0.740	0.833	0.858	1.731	0.321
Vanilla+R	5.75M	0.714	0.811	0.850	1.673	0.338	5.75M	0.757	0.842	0.860	1.799	0.311
Vanilla+R+L_{res}	5.75M	0.742	0.847	0.861	1.762	**0.282**	5.75M	0.776	0.851	0.866	1.875	0.299

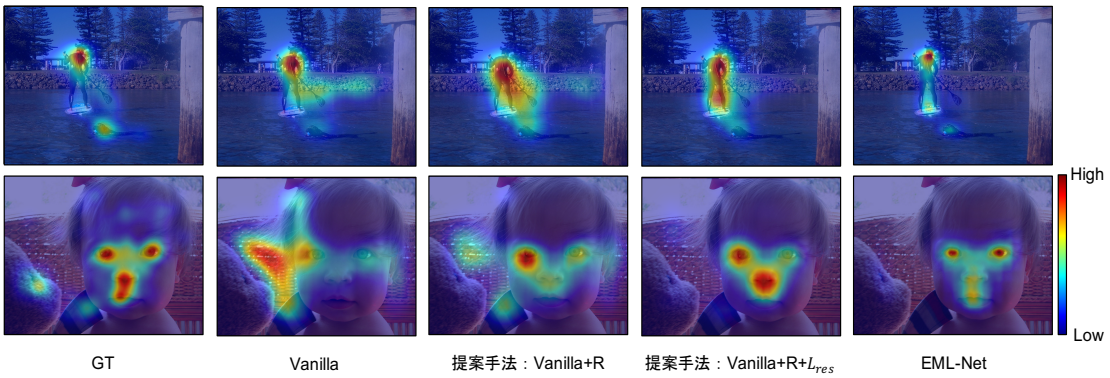

図 3：各手法による顕著性マップの推定例

ある．SAICON は学習用に 10,000 枚，評価用に 5,000 枚を，CAT2000 は学習用に 1,600 枚，評価用に 400 枚を用いる．また，学習モデルの汎化性能の観点から前処理として入力画像を [0,1] へ正規化，及び左右反転のデータ拡張を行っている．評価に利用する従来手法とモデルを以下のように定義する．

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：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(Ours)：Refinement layer 及び MixConv を活用して獲得した特徴を統合して Deconvolution 層へ入力するモデルである．

Vanilla + R + $\mathbf{L_{res}}$(Ours)：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．

4.2 ベースラインとの定量的・定性的な比較

顕著性予測における定量的評価指標として，SIM(Similarity)，CC(Correlation coefficient)，AUC(Area Under Curve)，NSS(Normalized Scanpath Saliency)，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とパラメータ数を表 1 に示す．提案手法は，整合性を保つ損失関数 L_{res} と Refinement layer により，従来手法と比較してほぼ同等の精度であることがわかる．また，IRL と MixConv を利用することにより，パラメータ数を 87.7%削減できていることが確認できる．

次に，図 3 に各手法による顕著性マップの推定例を示す．図 3 より，Vanilla では岩や木材で編まれた椅子などエッジ等の特徴が密集する位置に誤って顕著性が現れている．一方，提案手法は Refinement layer と L_{res} の導入により誤りの少ない顕著性マップとなっている．

4.3 損失関数の有効性の調査

図 4 に，異なる解像度間の画像を入力としたときの CC による精度変化を示す．なお，元の解像度を横軸の 1 とし，右へ移るにつれて解像度が 2 分の 1 ずつ低下している．図 4 から，一貫性損失を用いていない Vanilla+R (Without L_{res}) と比べ，Vanilla+R+L_{res} (With L_{res}) では低解像度でも精度は低下しない事が確認できる．人間の視線では，32 分の 1 までの解像度であれば一定の顕著性マップが得られることから，各解像度間の整合性を考慮した一貫性損失を利用することで，人間に近い視覚的特徴を学習できているといえる．

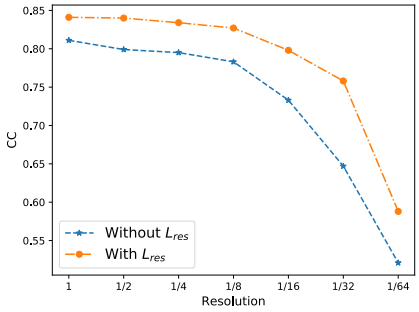

図 4：複数解像度に対する提案手法の精度比較

5. おわりに

本研究では，顕著性予測において MobileNetV2 や MixConv を利用したネットワークの設計によりパラメータ数を約 87.7%削減した．また，各解像度間の顕著性の整合性を保つ損失関数 L_{res} の提案により，L_{res} を利用しない場合と比べ精度が向上し，従来手法と比べてパラメータ数を削減したにも関わらず，ほぼ同等の精度となることを確認した．今後は，更なる精度向上を狙うためにアンサンブルモデルや AutoML によるネットワーク構造の最適化を行う予定である．

参考文献

[1] N. Riche, *et al.*, “A multi-scale rarity-based saliency detection with its comparative statistical analysis”, SPIC, 2013.

[2] J. Pan, *et al.*, “Shallow and Deep Convolutional Networks for Saliency Prediction”, CVPR, 2016.

[3] S. Jia, *et al.*, “EML-NET:An Expandable Multi-Layer NETwork for Saliency Prediction”, IVC, 2020.

[4] T. Judd, *et al.*, “Fixations on low-resolution images.”, JOV, 2011.

研究業績

[1] 瀬尾俊貴 等, “FlowNetC を導入した D&T における物体検出の高精度化”, 画像センシングシンポジウム, 2019.

[2] T. Seo, *et al*, “Video Object Detection and Tracking based on Angle Consistency between Motion and Flow”, IV, 2020.

(他 学会発表 1 件)