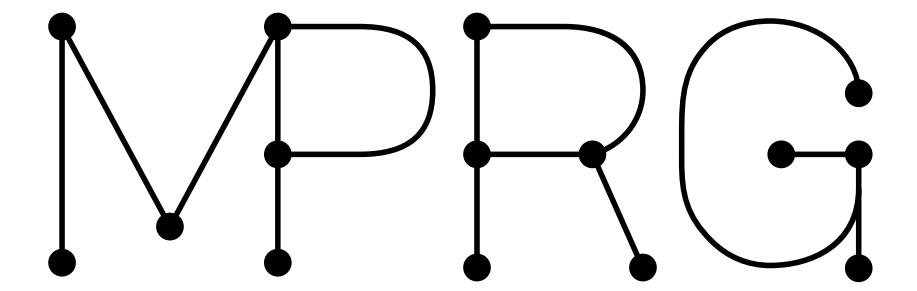


物体検出における注視領域の調査

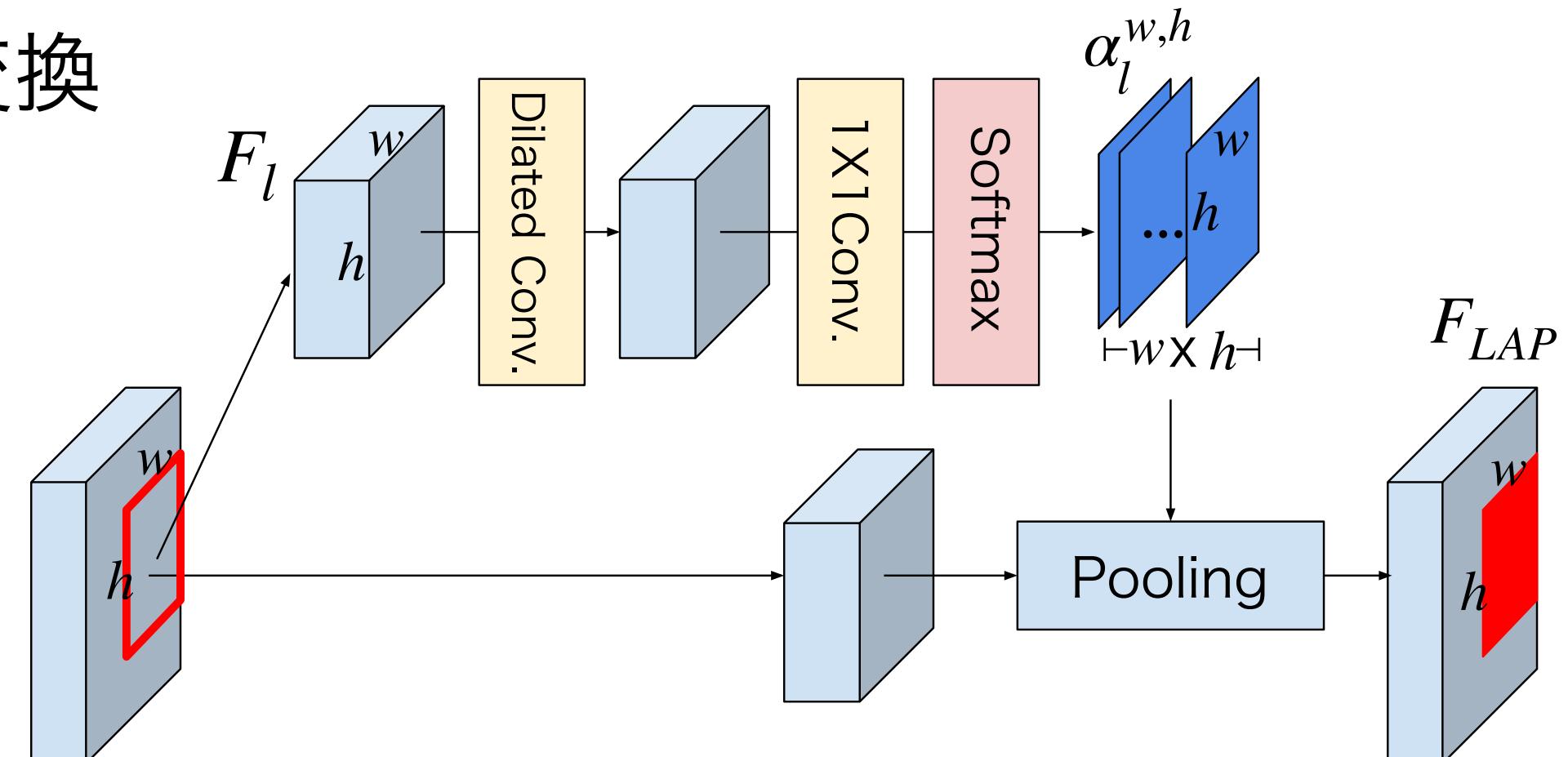
木村秋斗



MACHINE PERCEPTION AND ROBOTICS GROUP

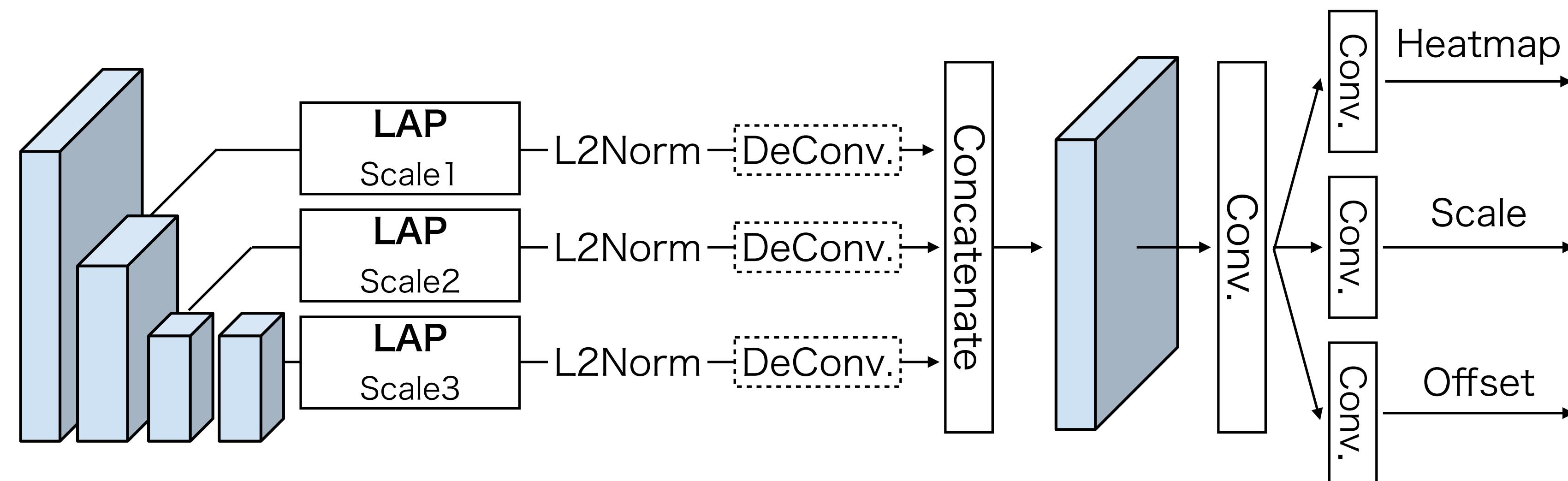
Local Attention Pooling (LAP) [N Liu+, CVPR2018]

- ・ 指定範囲の特徴マップ F_l から範囲内のコンテキスト情報のAttentionを含めた特徴マップ F_{LAP} を生成
- ・ Dilated convolutionによりコンテキスト情報を取得
 - 公開コードでは 7×7 の畳み込みを使用
- ・ 指定範囲の画素に対するAttention map $\alpha_l^{w,h}$ を取得
 - 1×1 の畳み込みでチャネル数を範囲内の総画素数に変換
 - 画素ごとの特徴をsoftmaxで変換
- ・ $\alpha_l^{w,h}$ と特徴マップ F_l をpoolingし F_{LAP} を生成



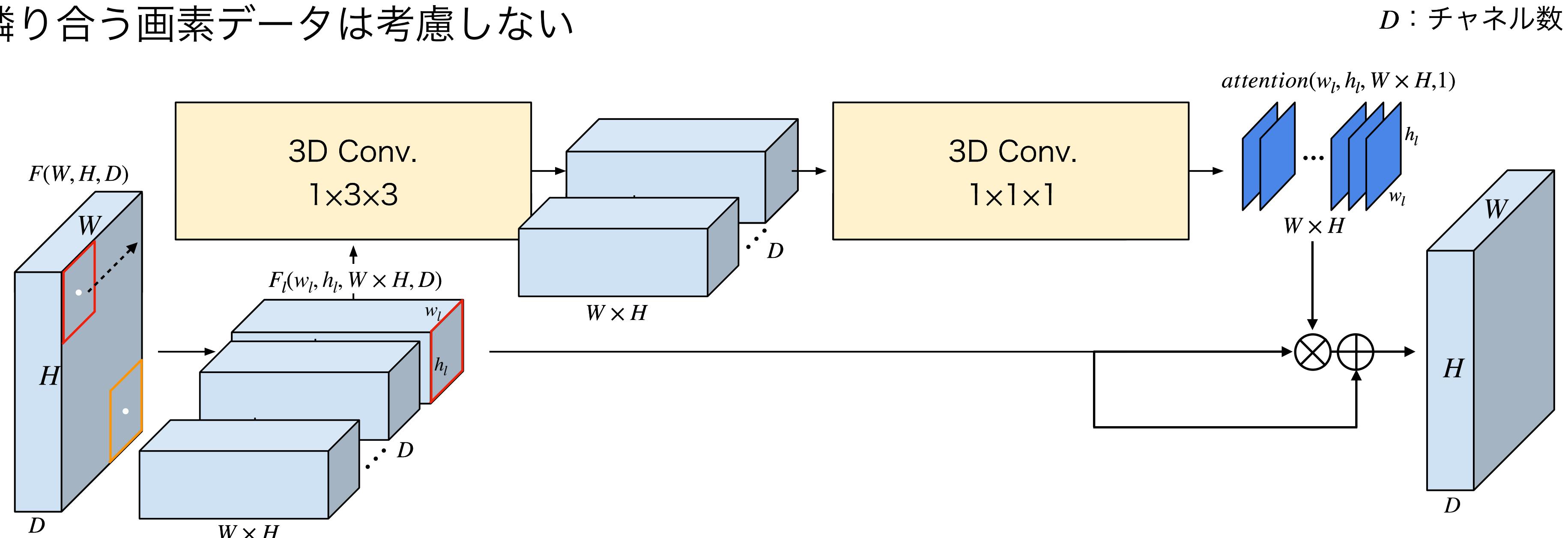
LAP構造をCSPへ導入

- CSPの各畳み込み層にLAP構造を導入
 - スケールごとに注視領域の変化があるかを調査



3D Convolutionによる導入案

- 対象の画素とその周囲をSliding windowの要領で順にスキャン
 - $w \times h$ (画素数分) の系列データを取得
- 3D Convolutionを用いて畳み込み
 - 画素数分のAttention mapを獲得
 - 時間方向のカーネルサイズは1で統一
 - 隣り合う画素データは考慮しない



実験結果 (同サイズ: 512×1024)

- ・ 学習画像サイズ
 - LAPなし: 512×1024
 - LAPあり: 512×1024
- ・ Smallの精度の差が小さい
 - 学習画像サイズの縮小が小さい歩行者の精度に影響
 - LAP導入による精度の低下はそこまで大きくはない

		MR[%]			
		Reasonable	Large	Middle	Small
LAPなし	Reasonable	20.2	23.2	5.8	15.1
	LAPあり	21.8	25.8	8.0	16.8

実験結果 (同サイズ: 1024×2048)

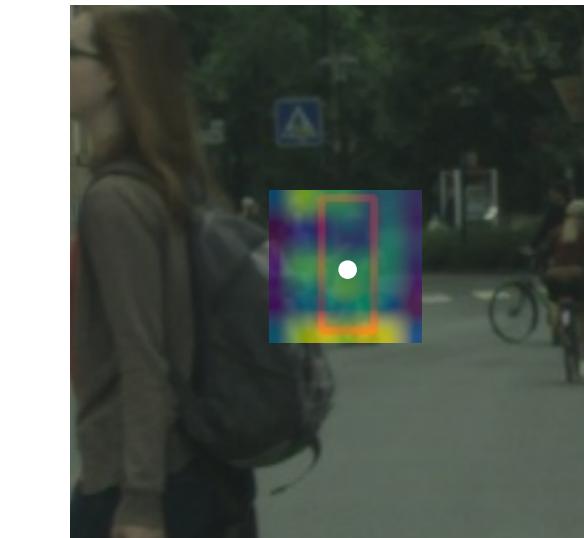
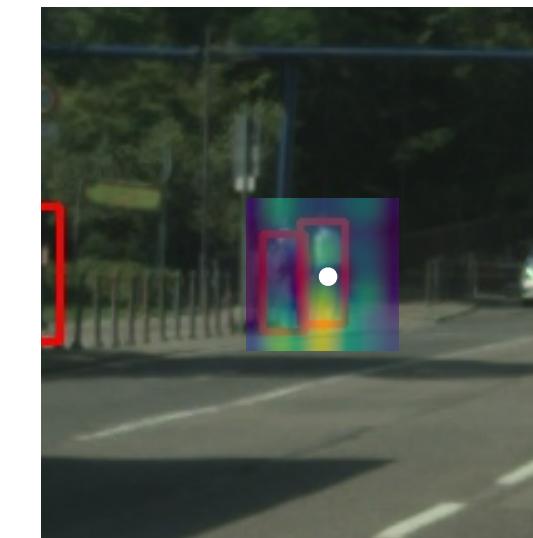
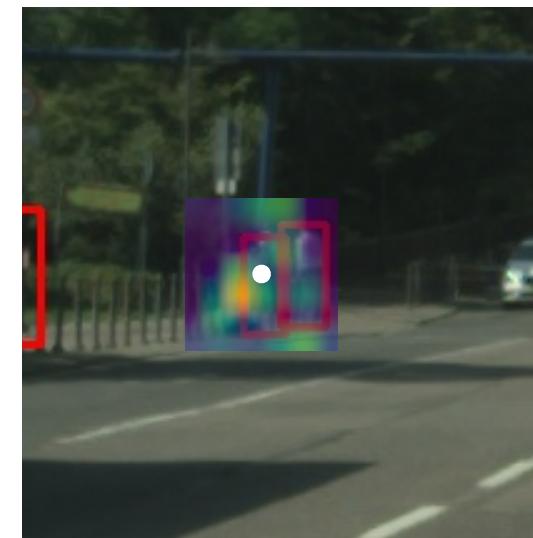
- ・ 学習画像サイズ
 - LAPなし: 1024×2048
 - LAPあり: 1024×2048
- ・ LAPの導入でReasonable, Largeの精度向上
 - LAPによる特徴マップの重み付けは画像サイズが大きい方が効果的

		MR[%]			
		Reasonable	Large	Middle	Small
LAPなし	Reasonable	15.8	24.7	6.4	7.3
	Large	14.7	21.3	6.8	7.8

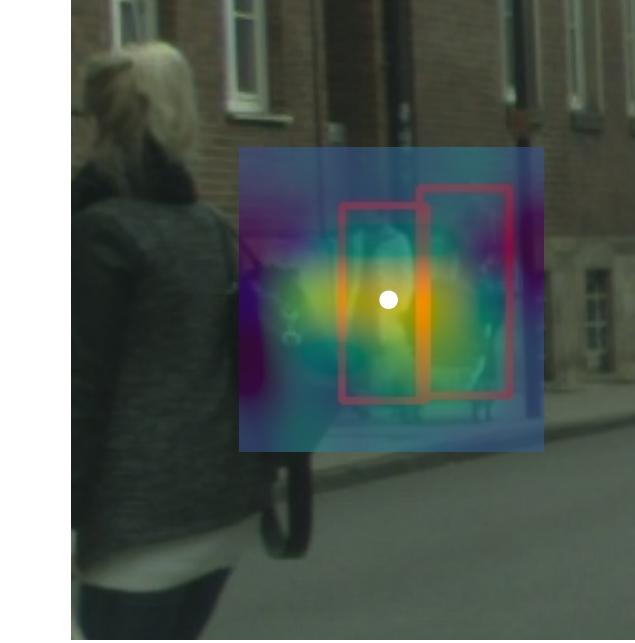
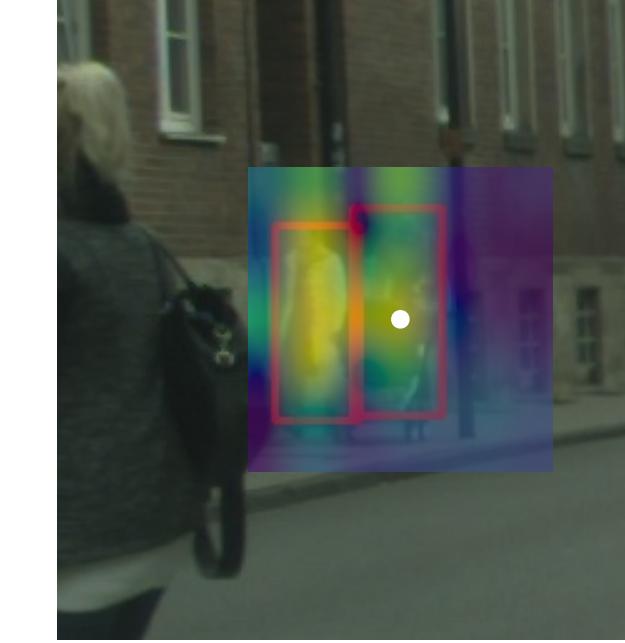
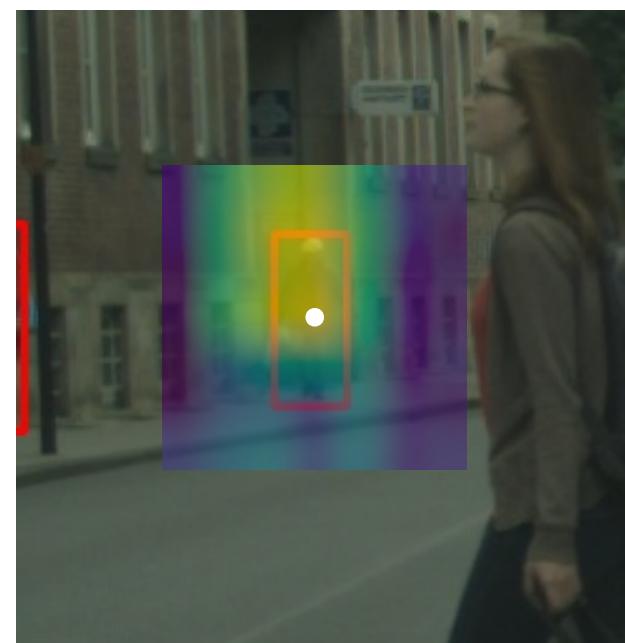
Attention mapの出力結果

- ・検出結果が歩行者全体を捉えた際の注視領域は上下端に対して強く注視する傾向

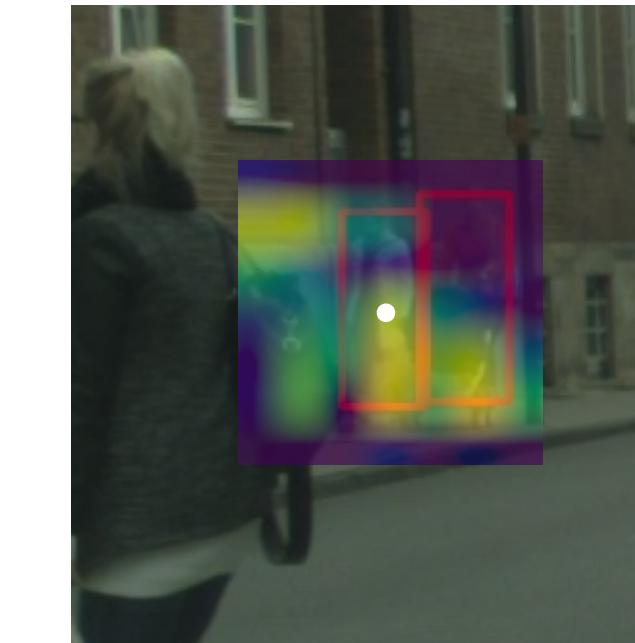
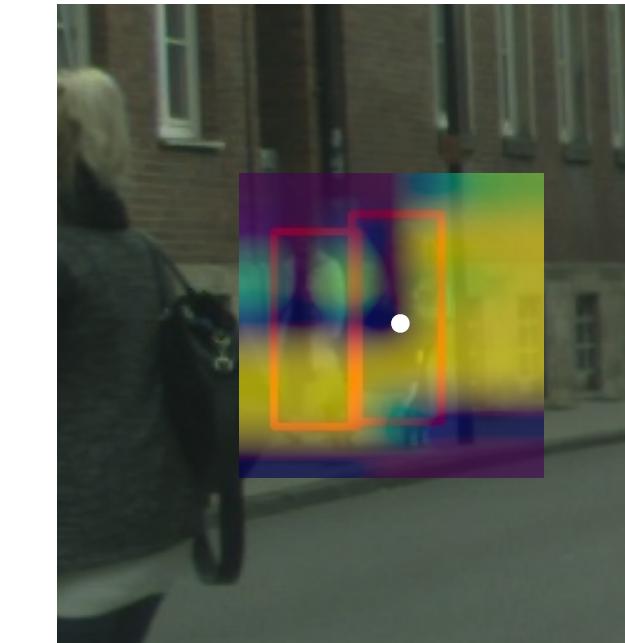
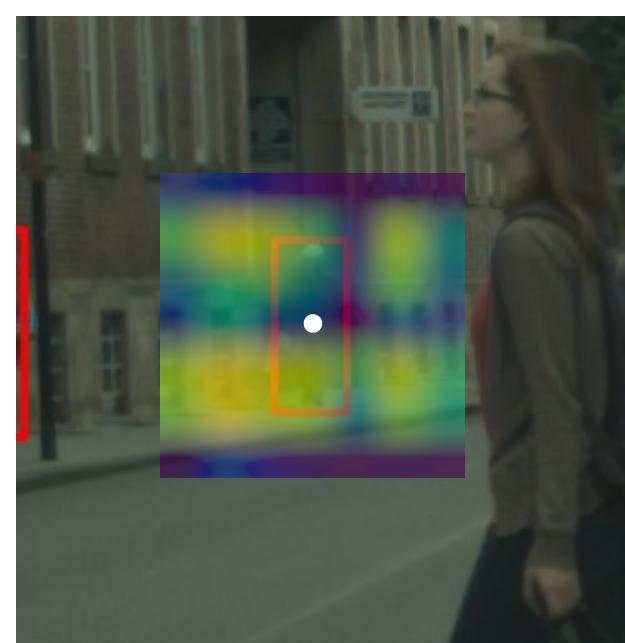
Scale1



Scale2

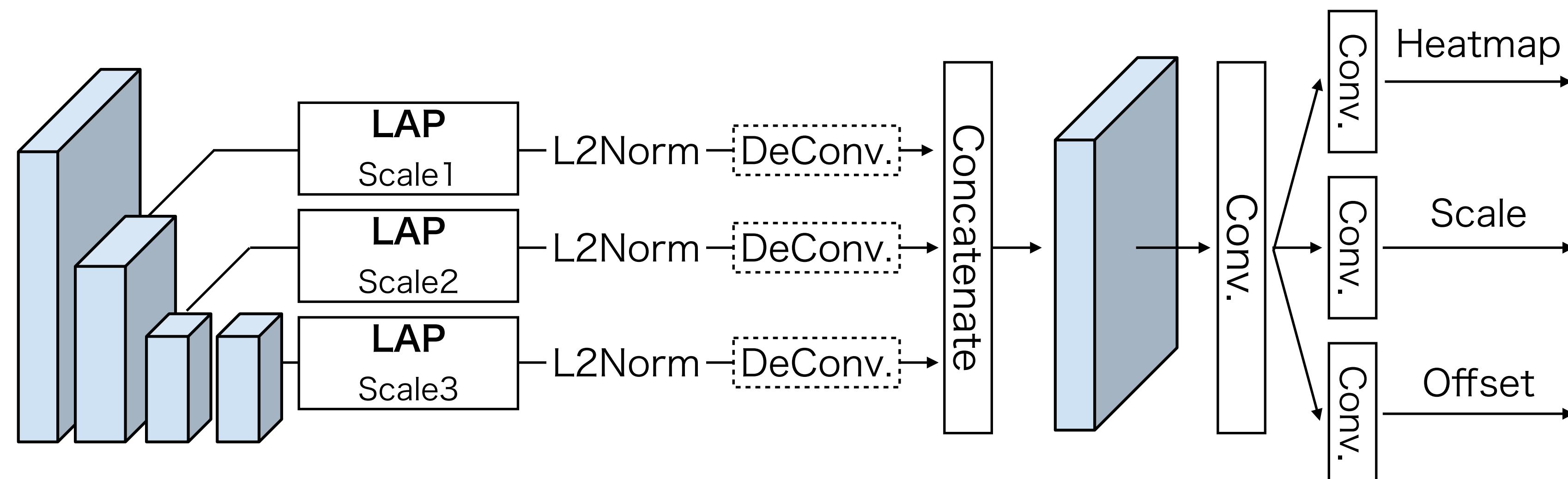


Scale3



学習対象の違いによる調査

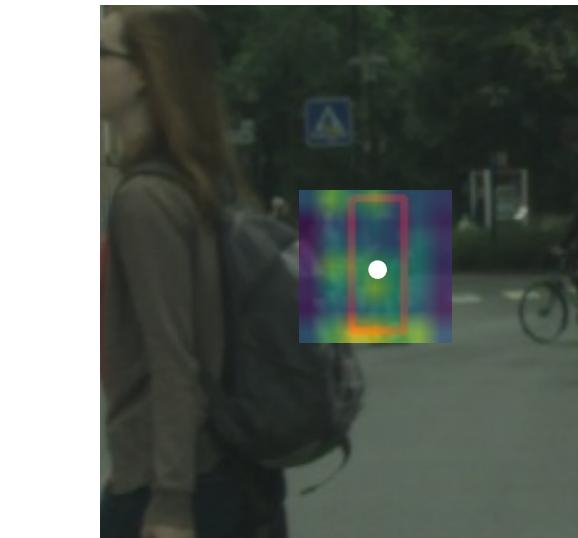
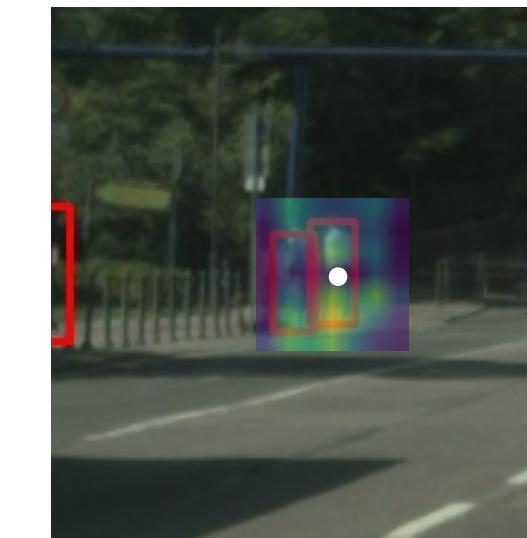
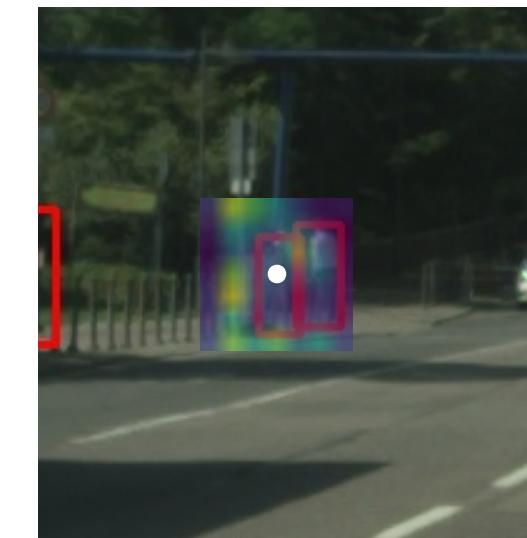
- CSPの学習対象を限定した際のAttention mapを調査
 - 中心の推定 (Heat map + Offset) のみにした場合
 - スケールの推定 (Scale) のみにした場合



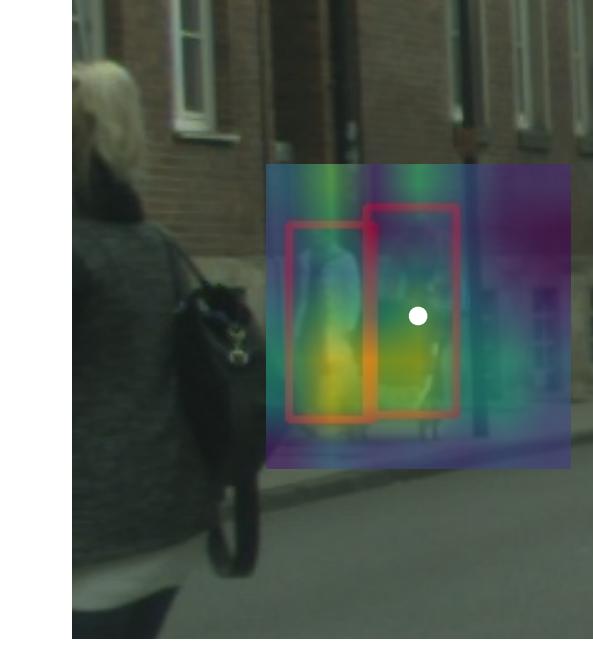
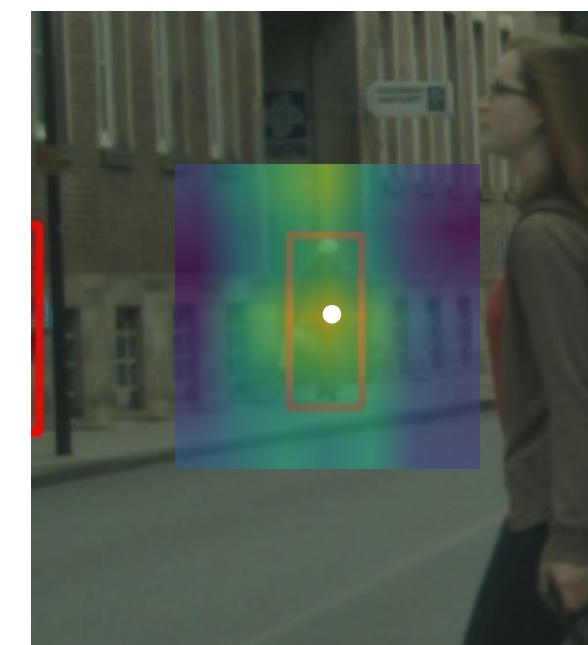
学習対象を中心点のみにした際のAttention map

- ・ 全体を注視するような傾向になる

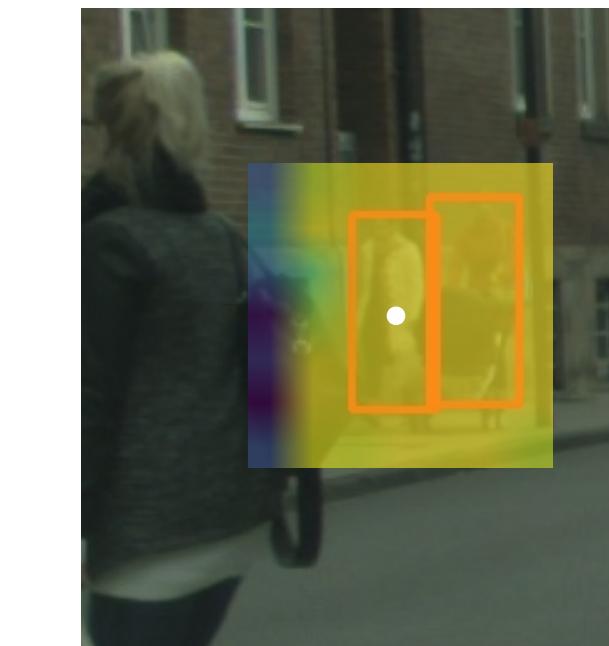
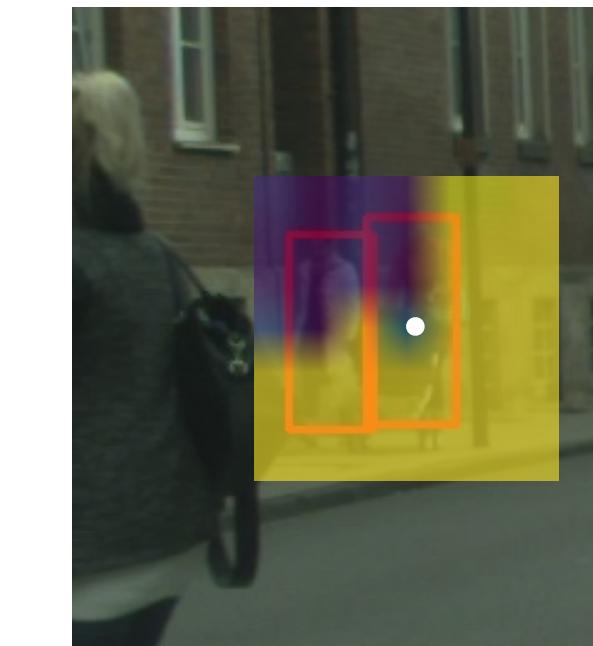
Scale1



Scale2



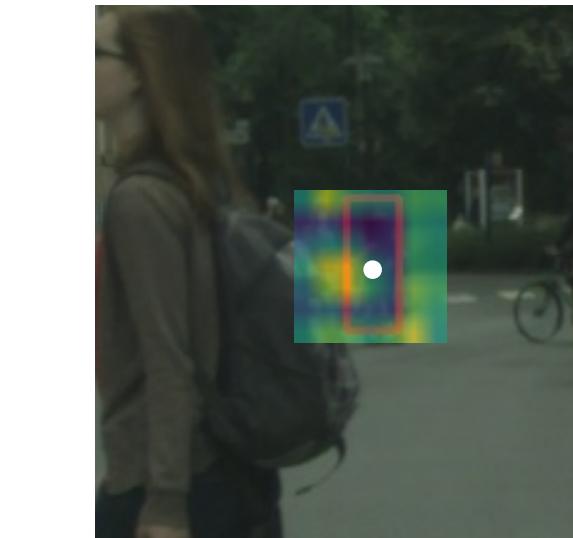
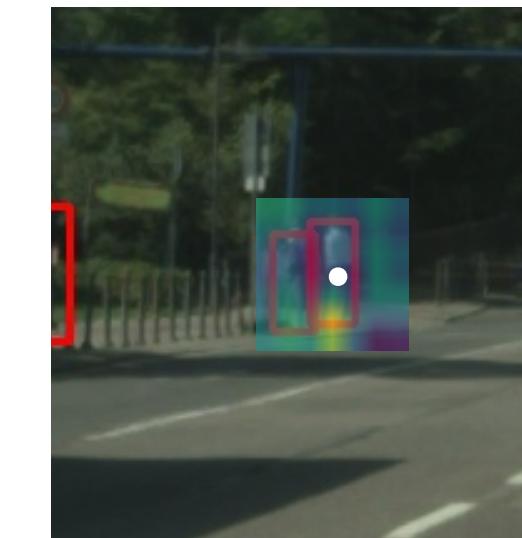
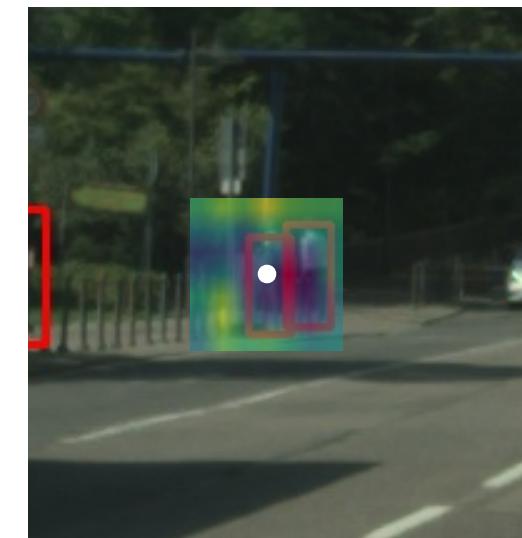
Scale3



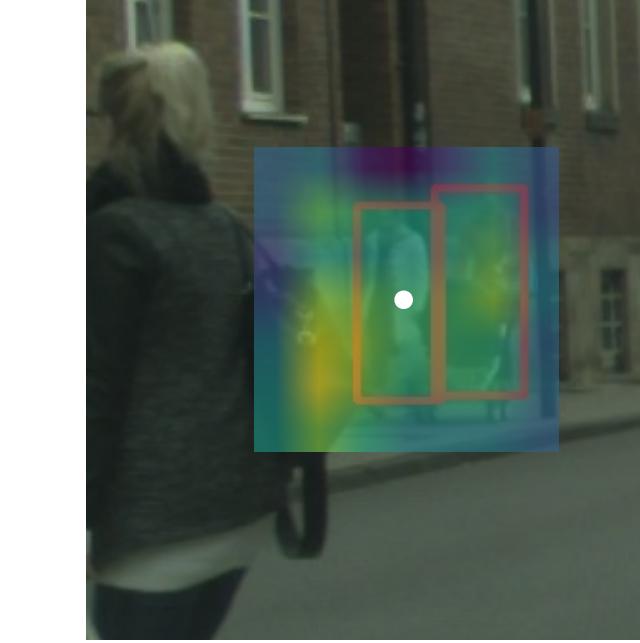
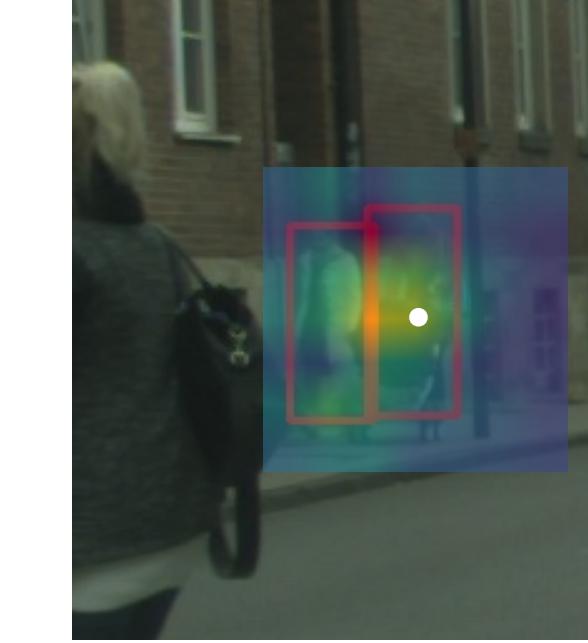
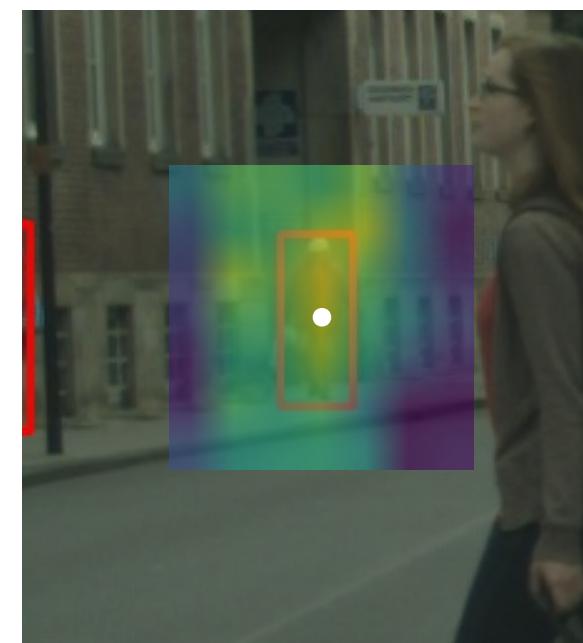
学習対象をスケールのみにした際のAttention map

- Scale2では曖昧な, Scale3では極端な注視領域を獲得

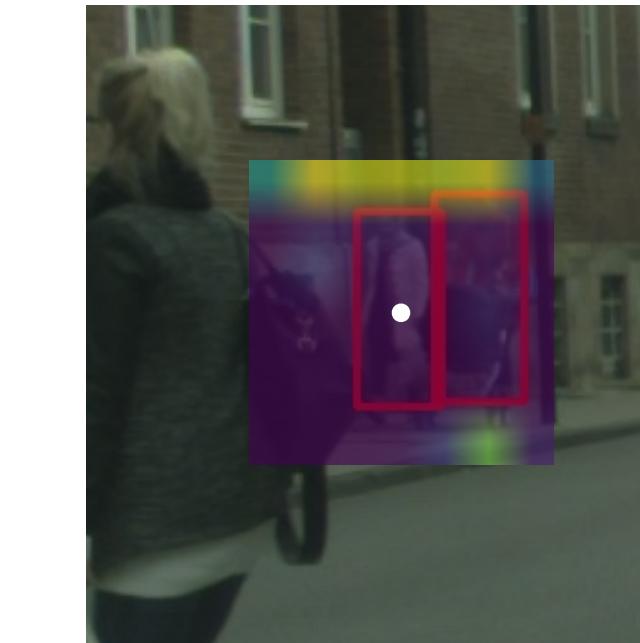
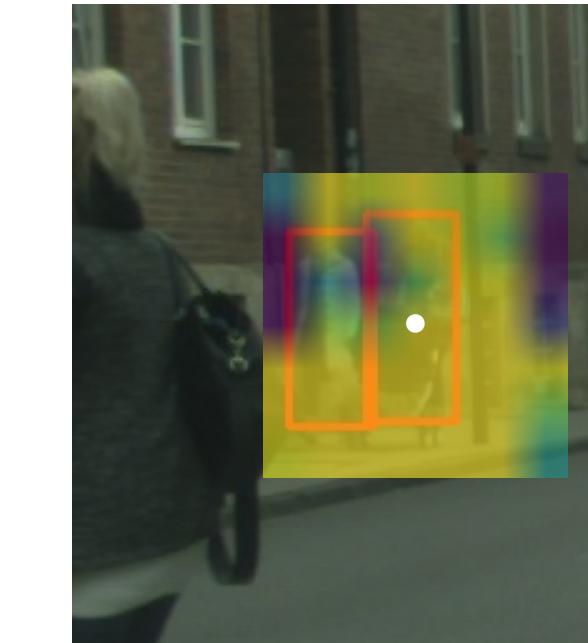
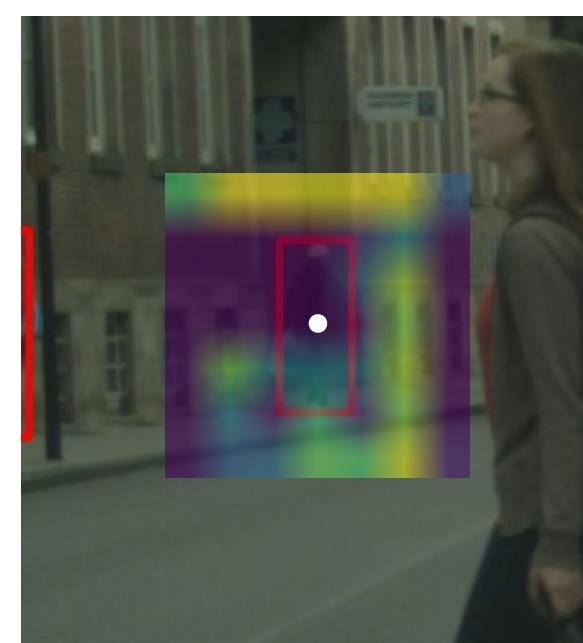
Scale1



Scale2

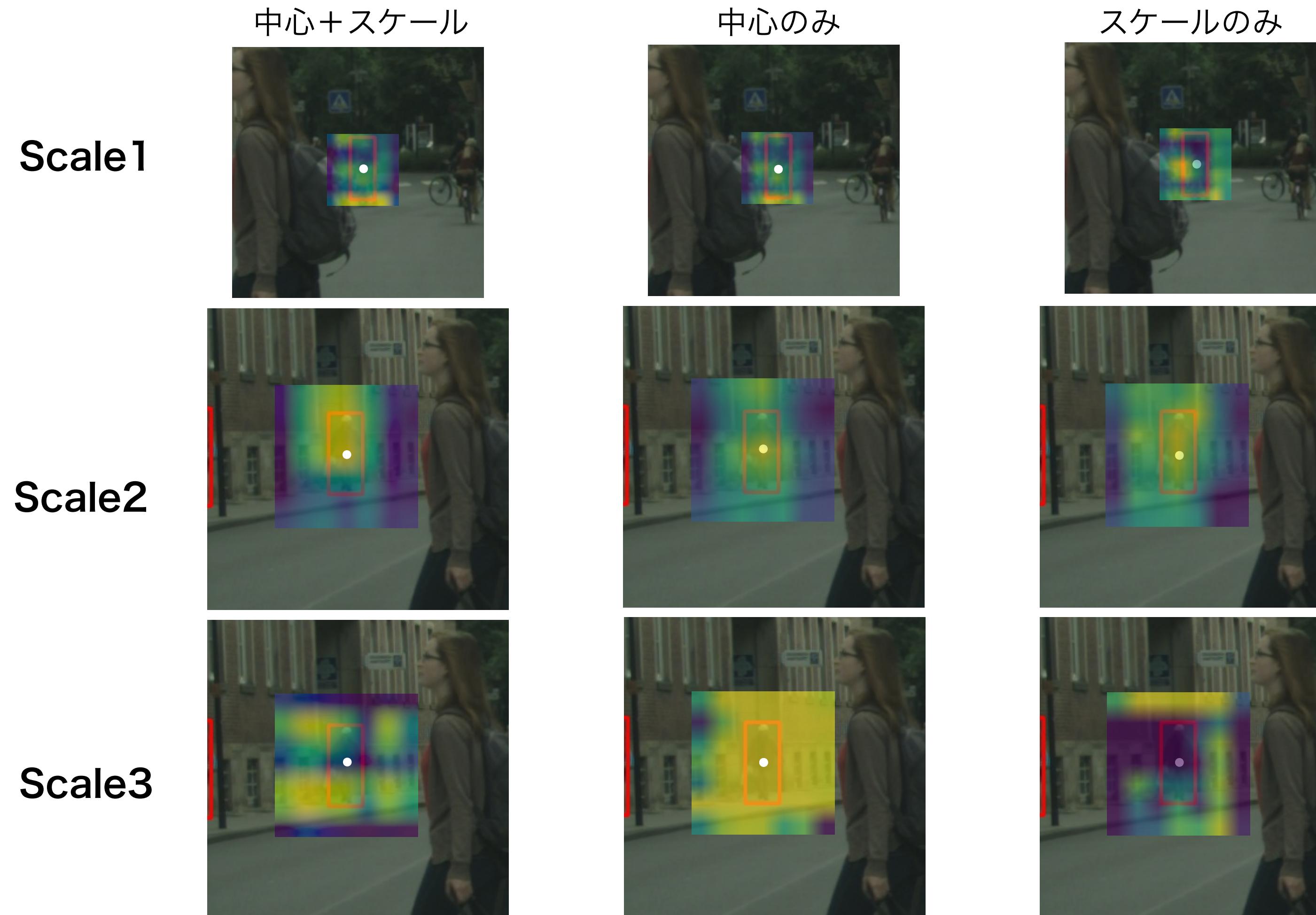


Scale3



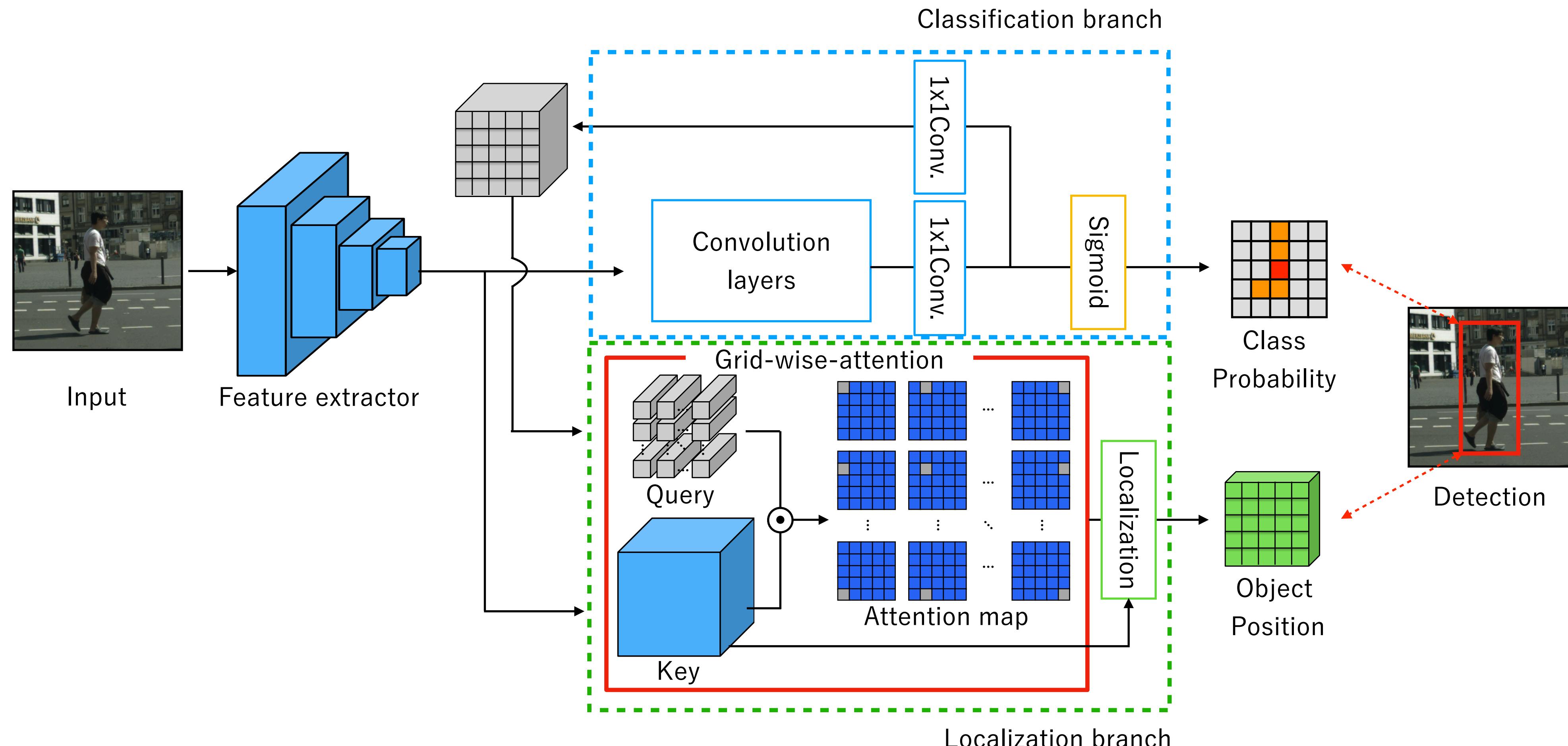
学習対象ごとのAttention mapの比較

- 中心+スケールと中心, スケール単独との相関はあまり見られない



grid-wise-attentionを導入した物体検出

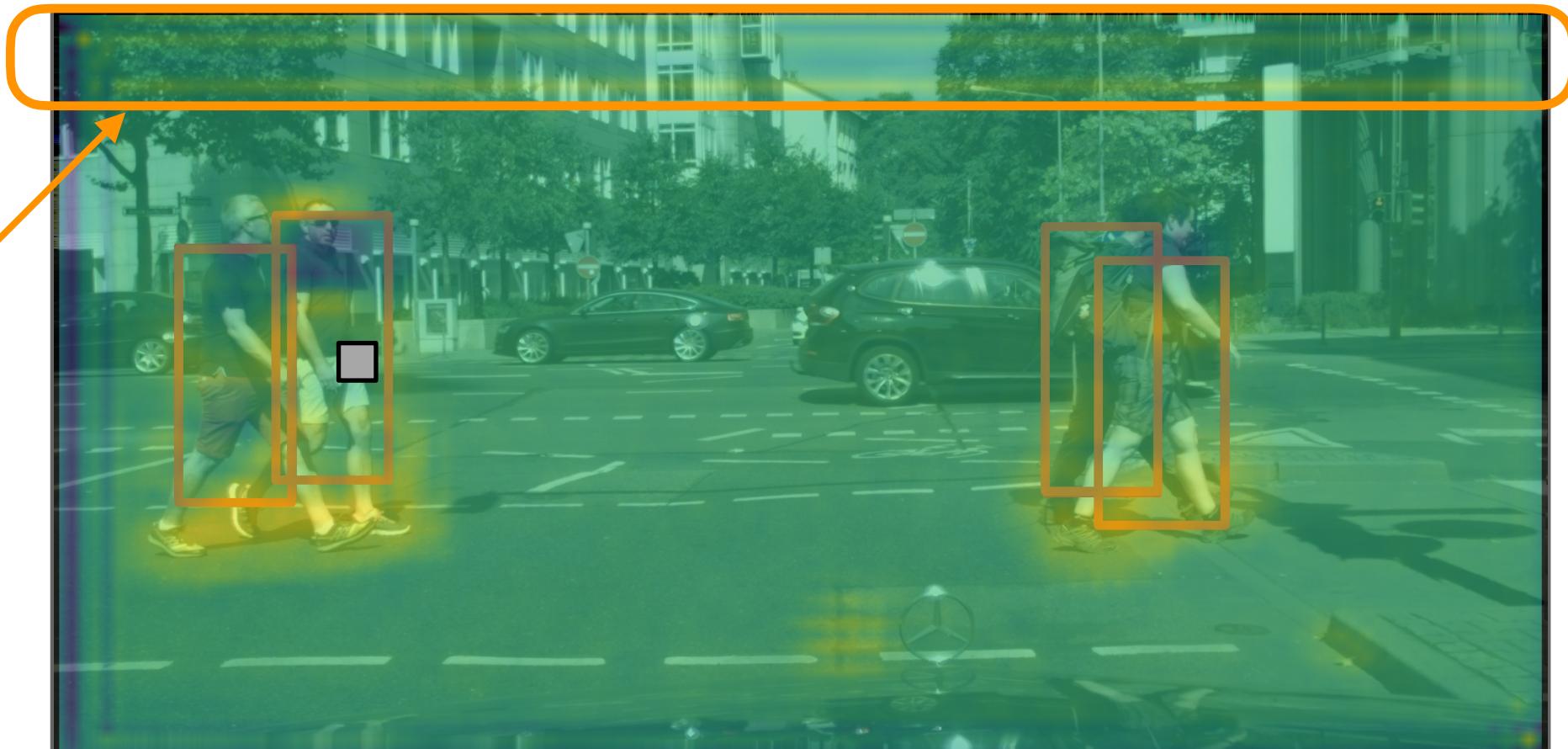
- ・ クラス分類と位置検出を分けて推定
- ・ 各グリッドに対する物体らしさを示すAttention mapを取得



- 従来のクラス分類のloss : Cross Entropy loss

$$L_{cls} = l_c(p_i, g_i)$$

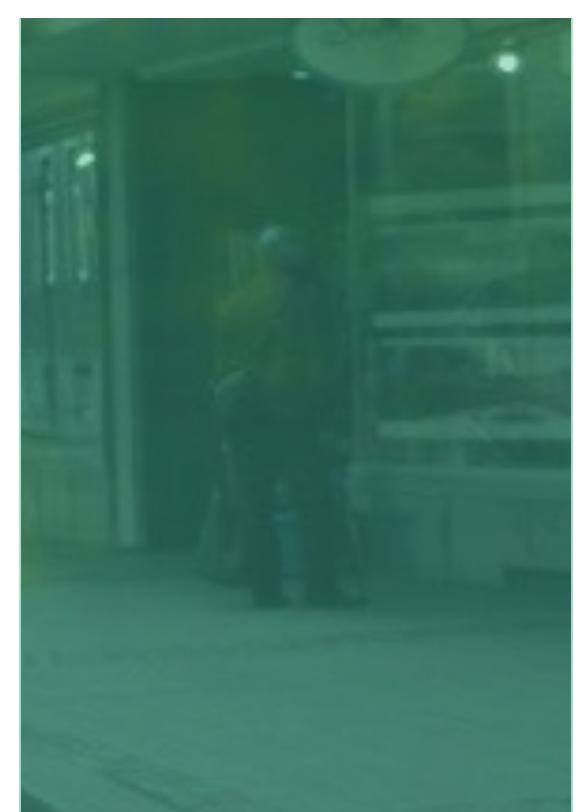
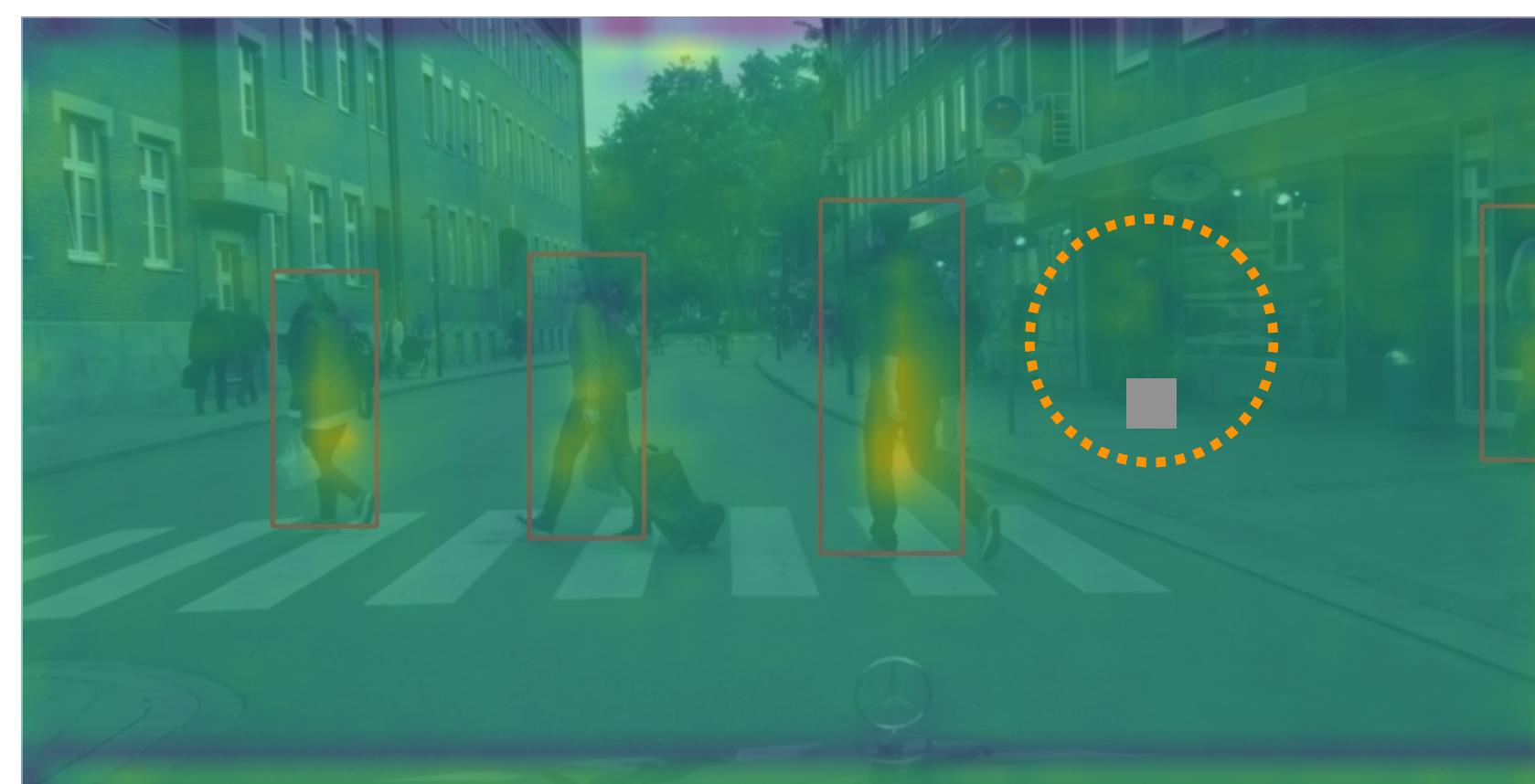
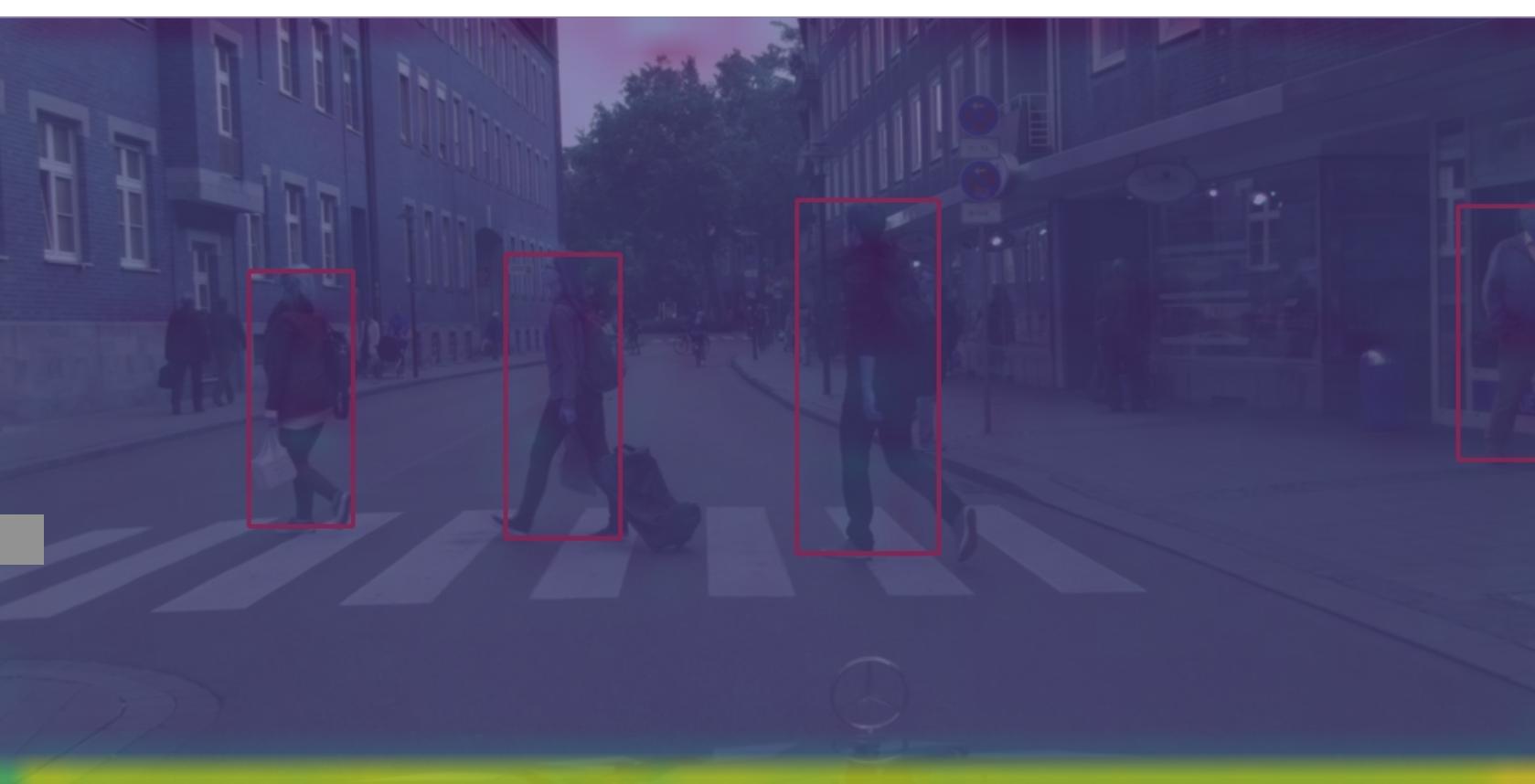
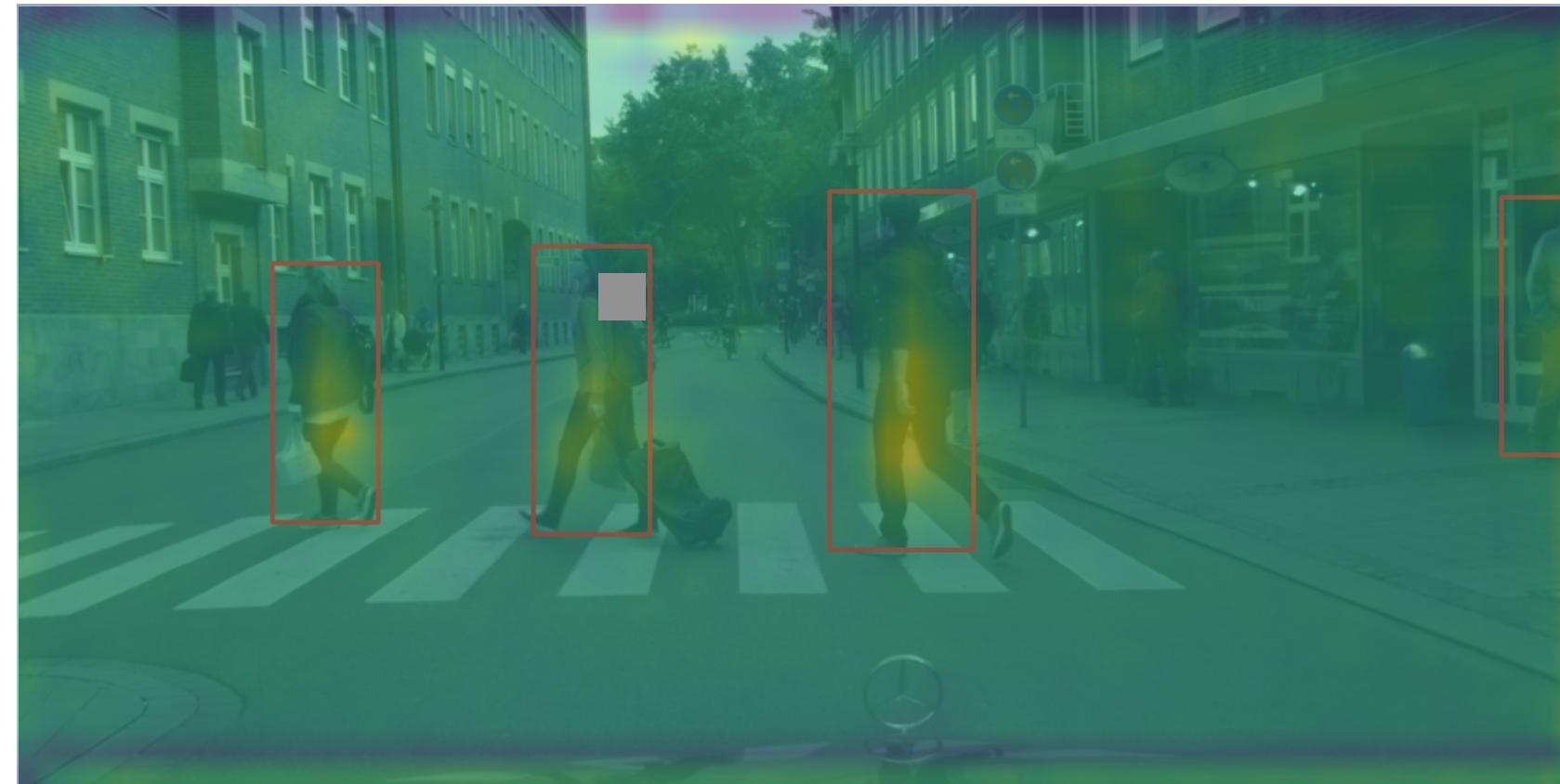
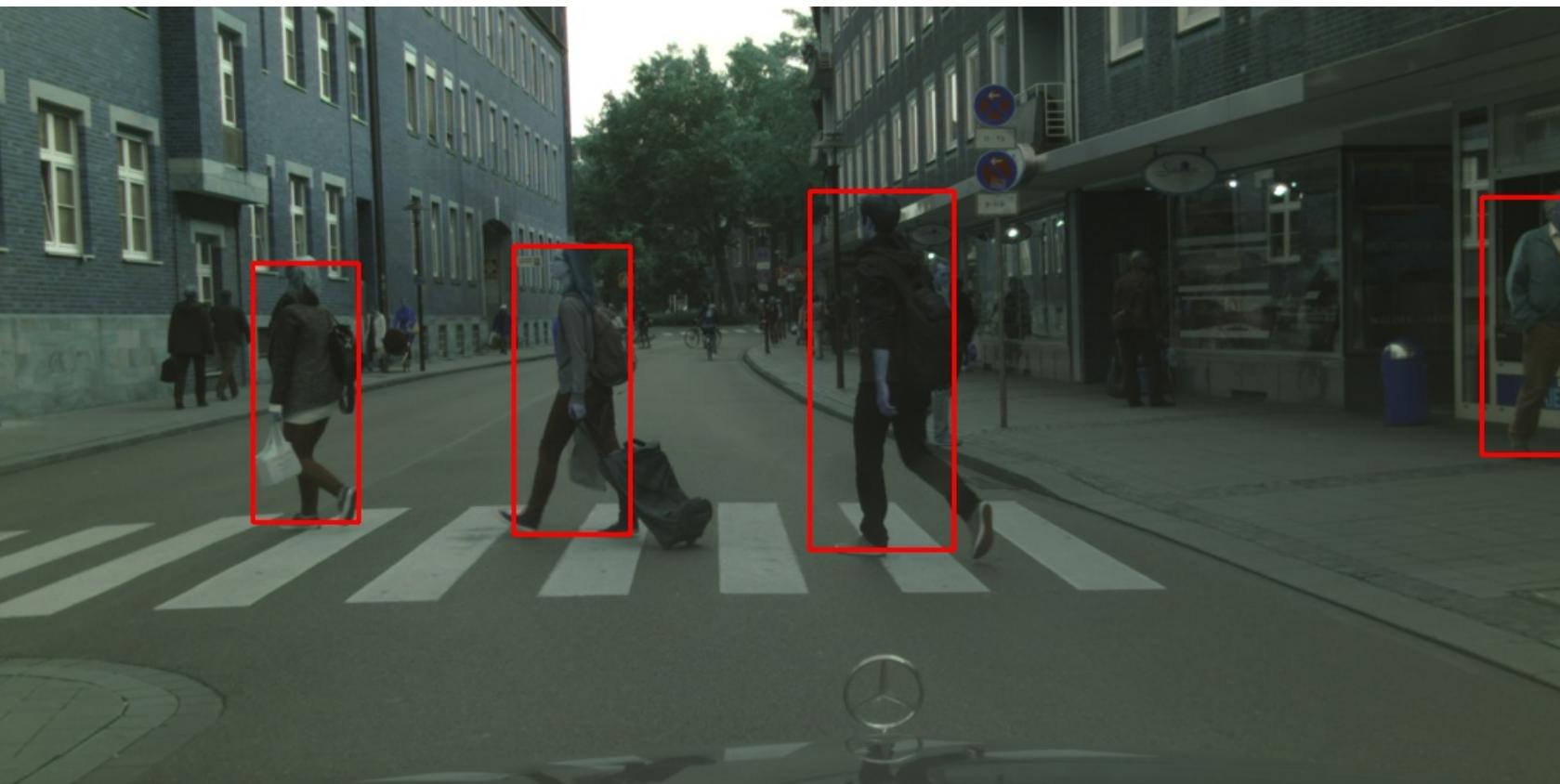
- 従来のAttention mapは背景の一部に強い注視
 - スコアは低いが背景の誤分類が発生？
- クラス分類のlossをFocal lossに変更
 - 背景（簡単で量の多い）分類に対するlossの重みを小さくするloss



$$L_{cls} = - \alpha_i (1 - \hat{p}_i)^\gamma \log(\hat{p}_i)$$

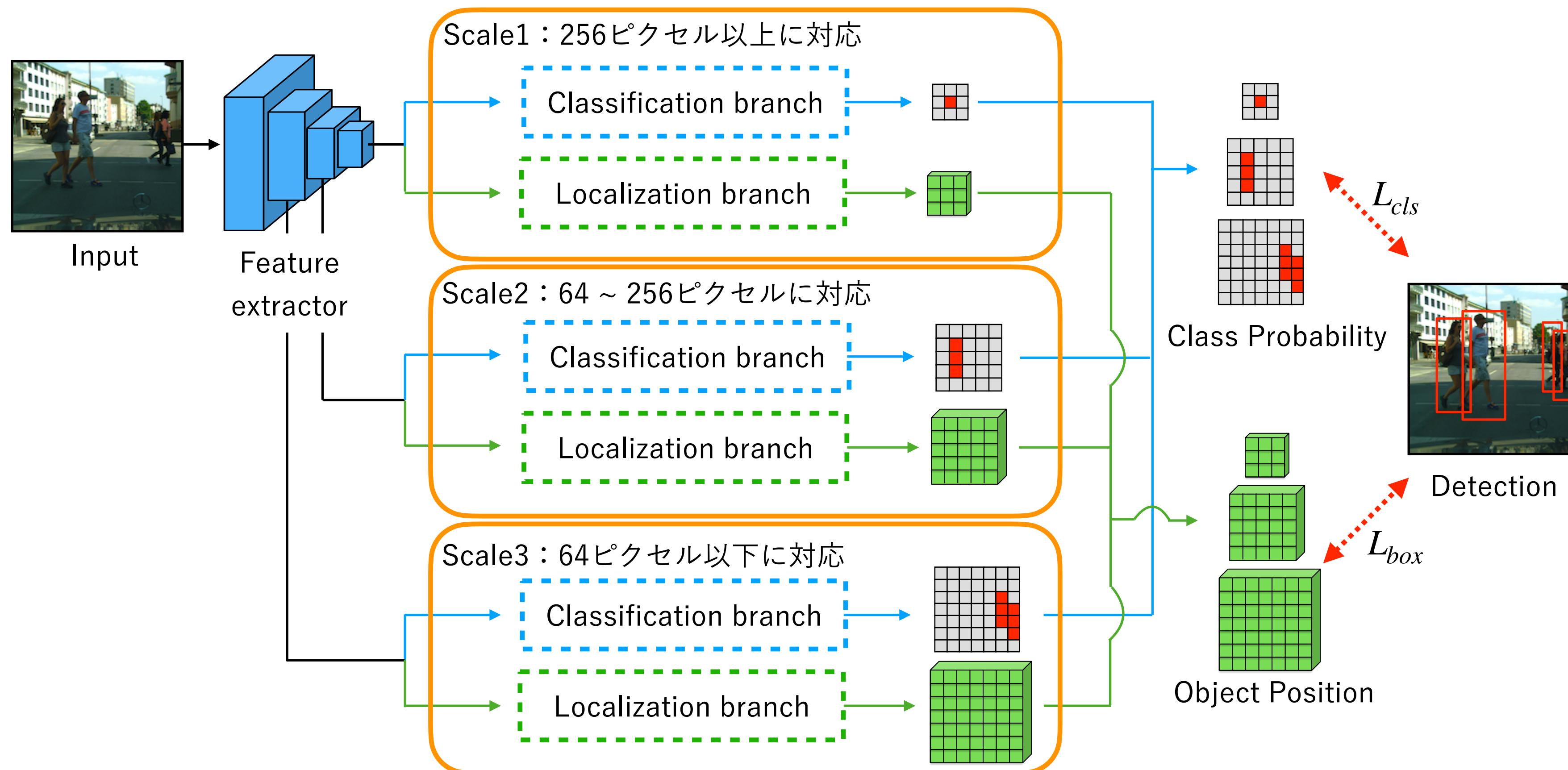
実験結果（定性的評価）

- 歩行者の体から足元にかけてを強く注視
 - 単一スケールのため、遠くの歩行者の検出や注視領域は出ない



マルチスケール検出

- 複数の異なるスケールで検出
 - より細かいグリッド分割で小物体の検出に対応



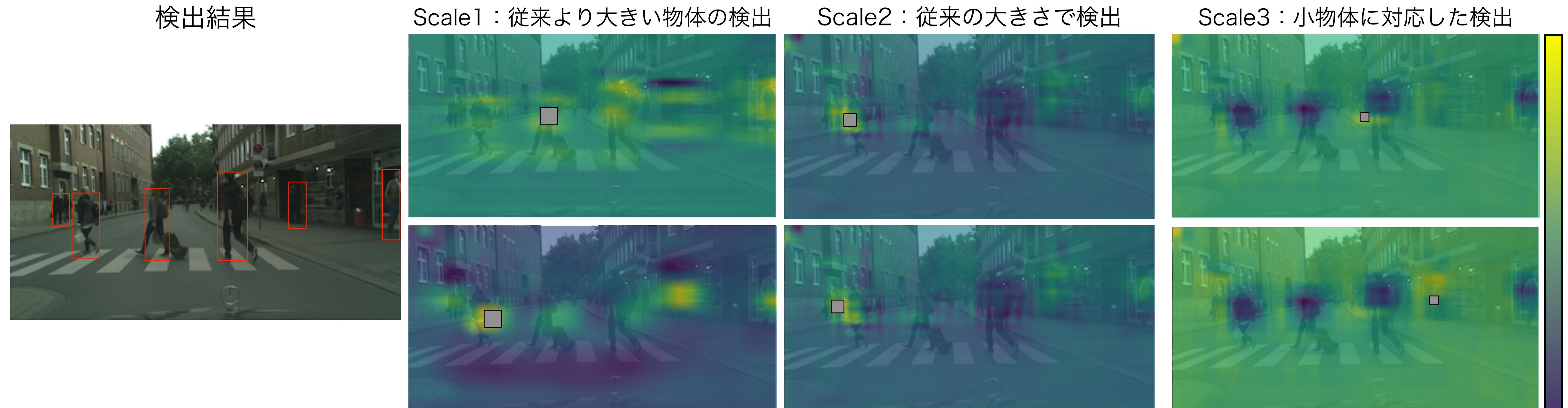
実験結果（定量的評価）

- Smallの精度がやや低い
 - 小物体の特徴を取得できていない

	All	Large (150 ~ pixel)	Middle (50 ~ 150 pixel)	Small (~ 50 pixel)
マルチスケール	18.1	18.4	6.1	31.2

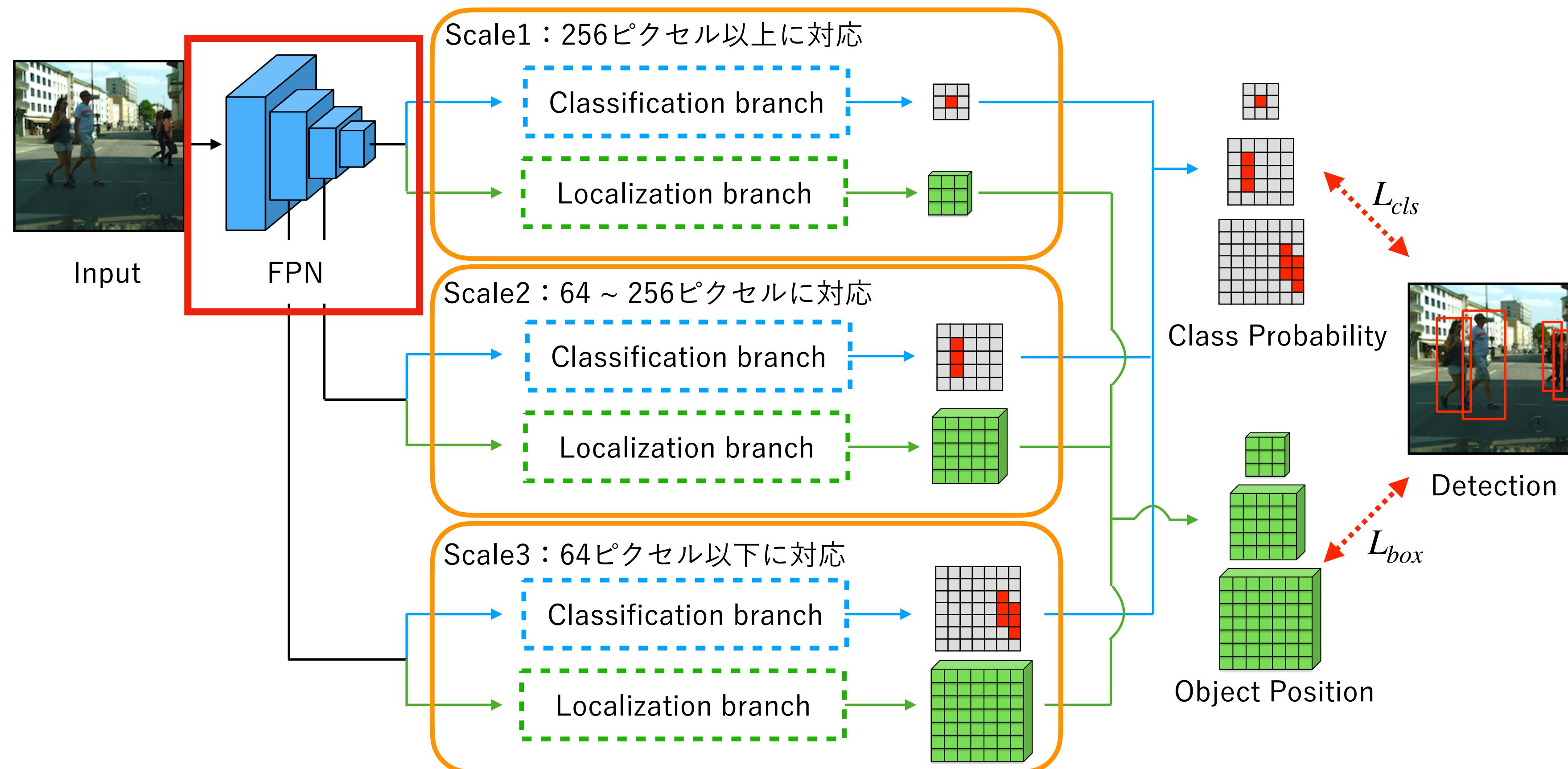
実験結果（定性的評価）

- ・歩行者の上半身付近を強く注視
- ・Scale3では大きい歩行者に対しては注視しない



ベースネットワークの変更

- ベースネットワークをFPNに変更
 - より高解像度で詳細な特徴を取得



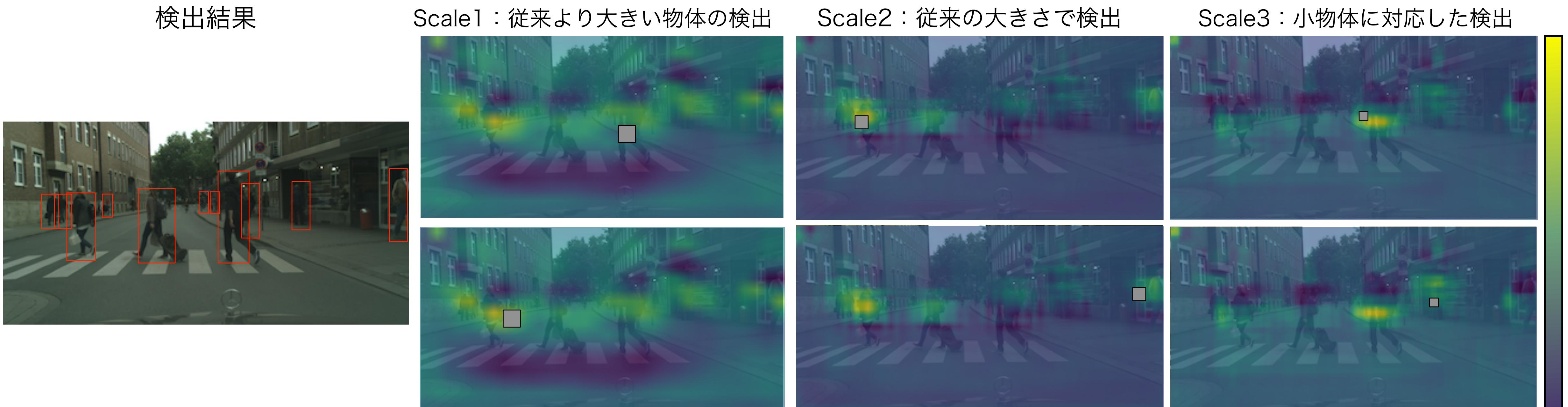
実験結果（定量的評価）

- ・ 全体的に精度が向上
 - 高解像度で詳細な特徴マップにより小物体の検出精度が向上

	All	Large (150 ~ pixel)	Middle (50 ~ 150 pixel)	Small (~ 50 pixel)
マルチスケール	18.1	18.4	6.1	31.2
FPN	13.1	10.8	6.3	20.0

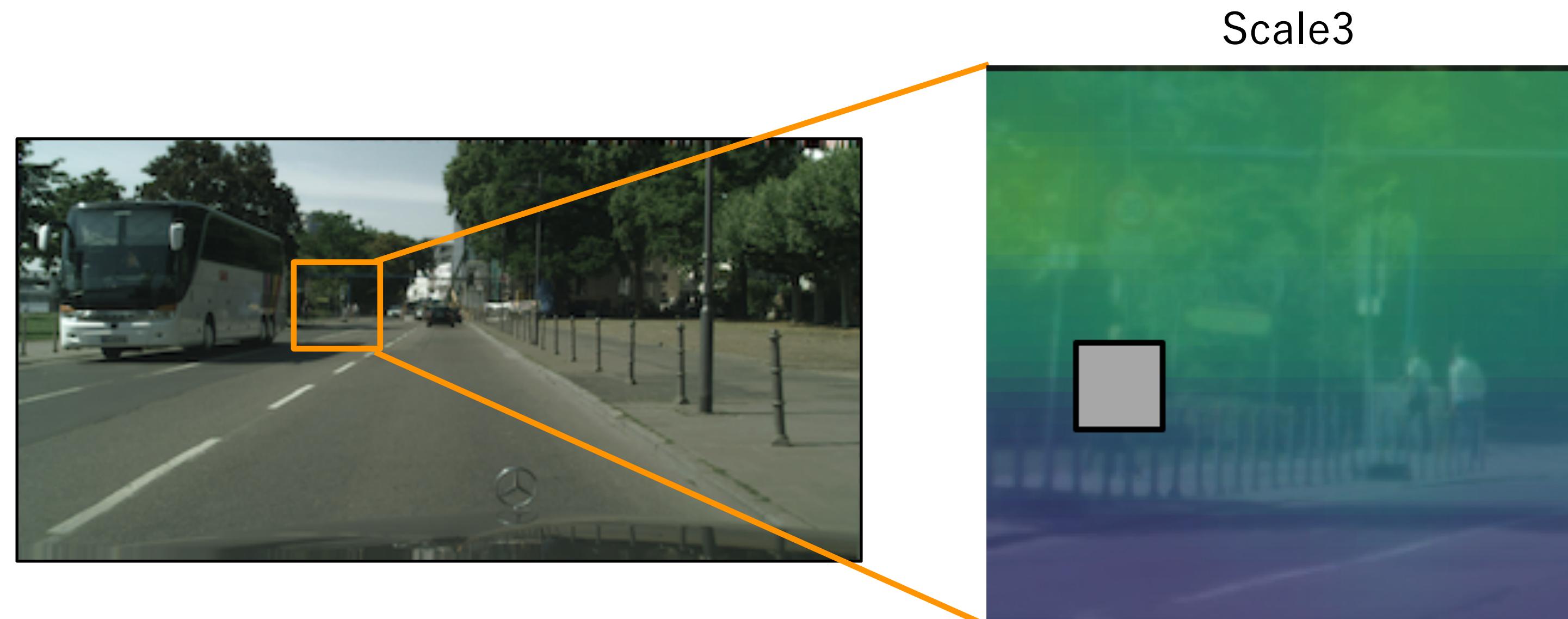
実験結果（定性的評価）

- 歩行者の上半身を強く注視
 - 頭付近には注視されない
- Scale3は大きい歩行者を注視



一部小物体に対する注視

- ・ 小さい歩行者が未検出
- ・ Scale3のアテンションマップが小さい歩行者を注視していない



今後の実装案

- Position encodingによってグリッドの位置情報を付与
 - gridに近い検出対象に対してより詳細に注視するアテンションマップの取得
- 検出方法をCenter and Scale Predictionに変更
 - 検出のパラメータ数を削減

- 本手法のアテンションマップ：検出対象以外の歩行者に対しても同じ傾向で注視
 - 重み付けが位置推定に影響
 - 対象物体により注視した方が高性能な位置推定が可能

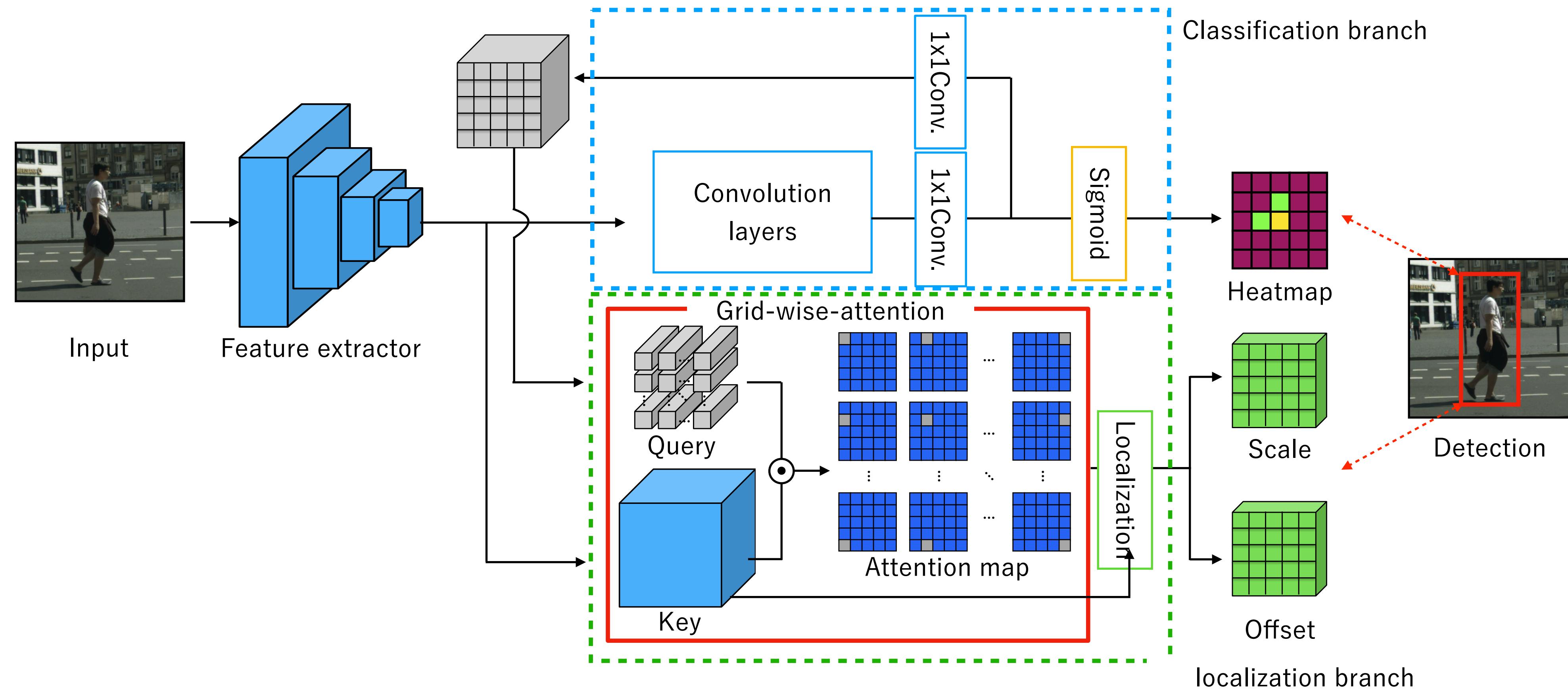


- 位置情報による注視領域の制限
 - 固定範囲のマスク：対象物体に具体的な注視
 - グリッドどうしの距離に応じた重みを付与
 - グリッド近辺の情報をより重視



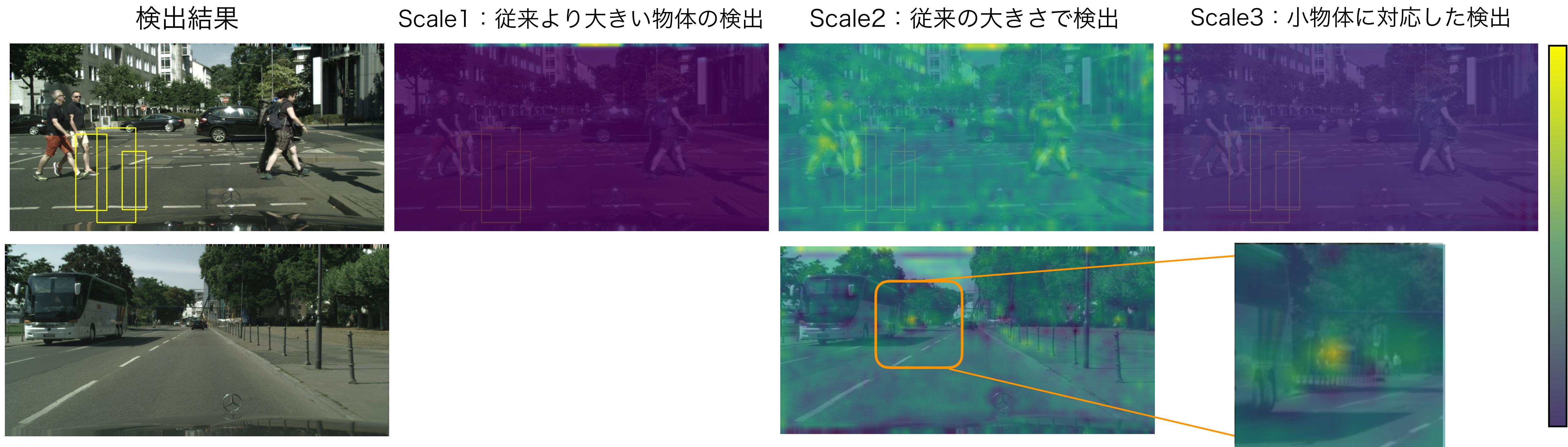
grid-wise-attentionを用いたCSPの検出

- Classification branchでHeatmap, Localization branchでScale, Offsetを出力
 - 1つのHeatmapで1つのクラスを推定する
 - より対象物体を考慮した特徴からAttention mapを取得



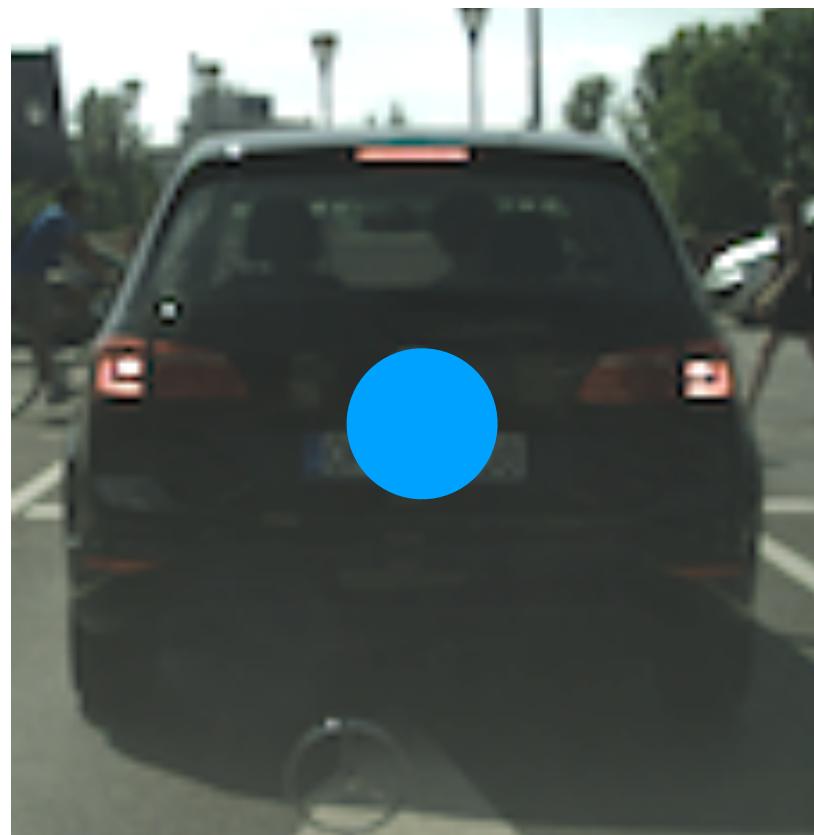
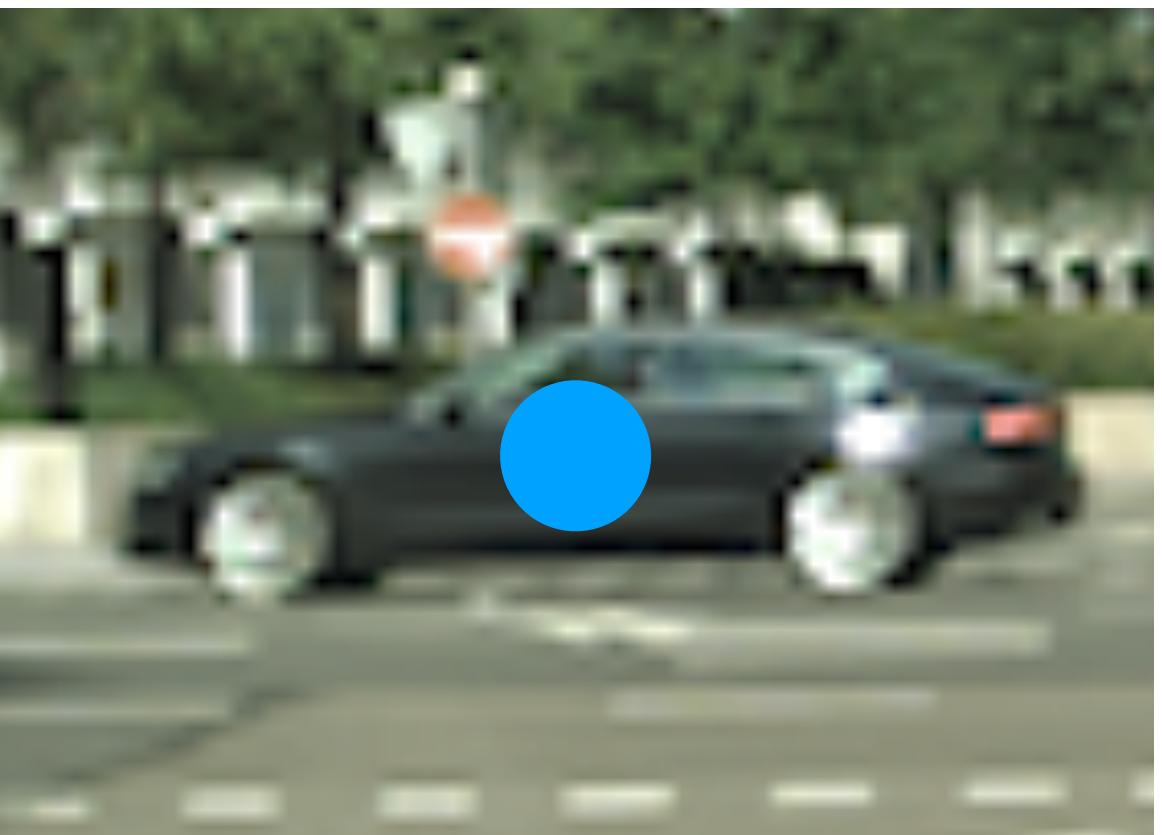
実装状況

- ・消費メモリ数が以前の約9割に減少
- ・35epoch終了時点の結果
 - 検出結果はまだうまく出ていないが、歩行者に対して注視されている
 - Scale2のみ歩行者に対して注視するアテンションマップを出力



CSPによる検出の問題点

- ・ アスペクト比が固定されていないと検出が困難
 - スケール推定のパラメータが増加
 - ・ 縦 → 縦と横
 - 同じ物体でも中心が変化しやすい

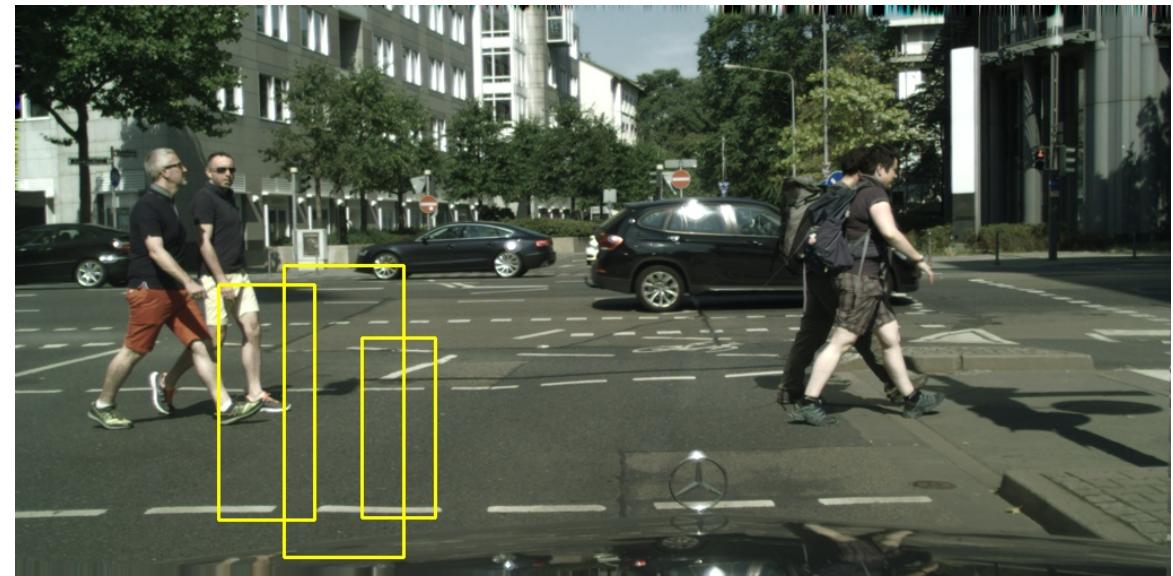


- grid-wise-attention
 - 損失関数の変更
 - マルチスケール検出
- 今後の予定
 - マルチスケール検出の精度向上
 - 位置情報の付与
 - gridに近い検出対象に対してより詳細に注視するAttention mapの取得

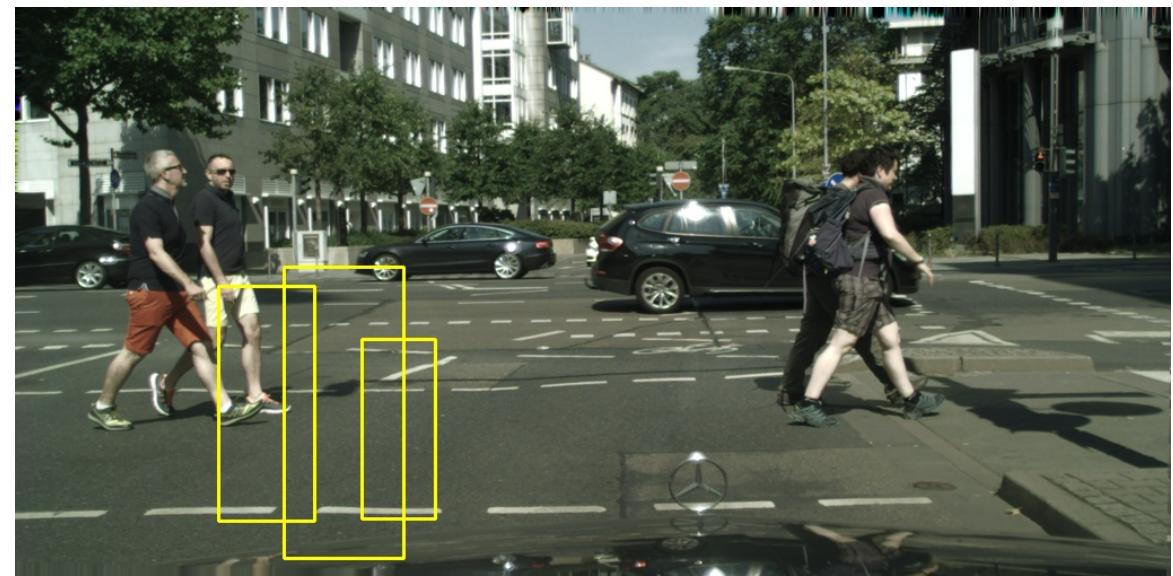
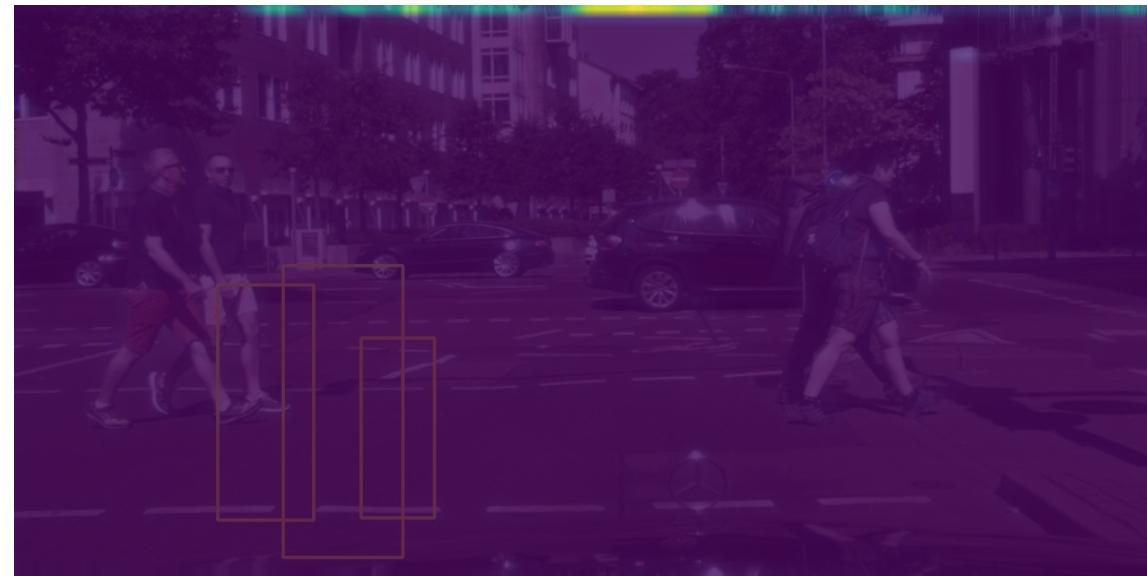
CSPによる検出のアテンションマップ

- Scale2のみ歩行者に対して注視するアテンションマップを取得

検出結果



Scale1：従来より大きい物体の検出



Scale2：従来の大きさで検出



Scale3：小物体に対応した検出

