DEIM2022 E34-3

作業記憶を考慮したタスク分割による セマンティックセグメンテーションタスクの効率化

小林　正樹†　　森田ひろみ††　　森嶋　厚行††

† 筑波大学 図書館情報メディア研究科　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail:　†makky@klis.tsukuba.ac.jp, ††morita@slis.tsukuba.ac.jp, ††morishima-office@ml.cc.tsukuba.ac.jp

あらまし　セマンティックセグメンテーションは，さまざまな分野への応用が可能な画像処理タスクであり，これまでにオープンベンチマークデータセットやそれを用いた深層学習モデルが提案されてきた．これらのデータセットでは，対象のドメインの専門家や匿名のクラウドワーカによってアノテーションが行われるが，高品質なデータセットの構築には多くの作業時間を要する．本論文では，人間の作業記憶の特性に基づいたタスク分割戦略により，画像へのセマンティックセグメンテーションタスクの効率化を目指す．実験結果から，ワーカ毎に 4 ～ 7 カテゴリの作業を割り当てることが，品質や総作業時間の観点から有効であることが示唆された．この結果は，人間の認知特性に基づくタスク分割により，ワーカからより良い作業結果を引き出し，かつ並列処理を可能にすることで，セマンティックセグメンテーションのためのデータセット構築の効率化に貢献するものである．

キーワード　クラウドソーシング, セマンティックセグメンテーション, 作業記憶

1　はじめに

画像へのセマンティックセグメンテーションは自動運転や工場の自動化などのさまざまな応用が挙げられるタスクである．これまでに道路を含む屋外や屋内を写した画像で構成されたデータセットやそれらを用いた深層学習モデルが提案されている [1] [2] [3] [4]．一般にセマンティックセグメンテーションのモデルの学習には大量の学習データが必要であり，例えば ADK20K データセットでは 25574 枚の画像に対してアノテーションが付与されている [5]．

セマンティックセグメンテーションを新たな課題に適用するにあたり，学習データセットを構築するためのアノテーションの効率化が重要である．データセットを作るためにはコストがかかることが知られており，Cityscapes データセットの提案の論文では 1 画像あたり 90 分を費やして専門家によるアノテーションが行われたと報告された [1]．十分な品質と量の学習データを入手することができなければ，新しいドメインでの学習モデルの構築や評価が難しくなる．

一方で，論文等で言及されているアノテーションの方法では専門のアノテータを雇用して作業を依頼する方法か，クラウドソーシングを伴うワークフローを構築してアノテーションを行う方法が用いられている．専門家とは限らないクラウドワーカなどに作業を依頼する場合，品質管理などの観点からタスクを細分化が求められる．一部のアノテーションの効率化に関する研究では，ML モデル等を用いてアノテーションが必要な画像や領域を削減する方法が削減されている [6] [7]．この方法では，人間のワーカの作業量を減らすことができるが，一方で適用可能な学習モデル等が制限される．また，学習モデル等の評価のためには画像の全ピクセルへのアノテーションが必要である．

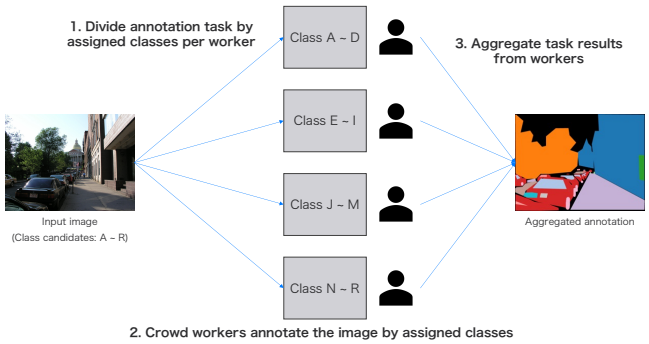

図 1　本研究では，人間の作業記憶に基づくタスク分割によるセマンティックセグメンテーションタスクの効率化手法を提案する．

本研究では，セマンティックセグメンテーションのアノテーション作業を細分化することで，データセット構築の効率化する手法を提案する（図 1）．提案手法では，アノテーション作業を道路や人のような対象のカテゴリごとにタスク分割することで，1 タスクあたりの作業量を削減し，また並列での作業を可能にする．一方で，1 名のワーカに依頼するカテゴリ数の合理的な決定方法は自明ではない．そこで，本研究では，認知科学の分野における作業記憶（ワーキングメモリ）の特性に基づいてカテゴリ数を決定することが作業品質や作業時間の観点で有用であると仮定し，これに基づくタスク分割を行う．提案手法は特定の深層学習モデル等に依存しないタスク分割手法であり，対象の画像の全ピクセルへのアノテーションを入手できるので，これにより得られたデータセットからさまざまなモデル等を構築・評価することが可能である．また，他の効率化手法と組み合わせて利用することも可能である．

評価実験では，クラウドワーカに割り当てるカテゴリ数を比較する実験により，提案手法に基づくタスク分割の有効性を確認した．

2 関連研究

セマンティックセグメンテーションの大規模なベンチマークデータセットが開発されている．代表的なものとして Cityscapes [1], Microsoft COCO [2], ADE20K [5] が挙げられる．また，深層学習に基づくセマンティックセグメンテーションのモデルとして，U-Net [3] や SegNet [4] が挙げられる．

深層学習モデル等をアノテーション作業に組み込むことで，作業者の負荷を削減したり，最終的に得られる学習済みモデルの性能を向上させる研究がある．ScribbleSup [6] は，作業者に画像の全てのピクセルにラベルを付与する代わりに，例えば車両の領域内に線を描いてもらい，その情報を周囲のピクセルに伝播させることで，不完全なアノテーションからセマンティックセグメンテーションのモデルを学習する手法を提案した．より作業者からの入力を削減できる手法として，画像に対する単一のラベルを入力とし，セマンティックセグメンテーションモデルを反復的に学習することで学習する手法が提案されている [8]．Lin [7] らは，画像へのセマンティックセグメンテーションタスクを画像を複数のブロックに分割し，一部のブロックのみへのアノテーションを作業者に依頼し，得られたアノテーションに基づいて学習したモデルにより残りの画像のセグメンテーションを行う手法を提案した．本研究の提案手法は，特定の入出力を備えた深層学習モデル依存せず，またアノテータに画像の全ピクセルへの作業を依頼するものである．そのため得られるデータセットは様々なモデルの学習及び評価に利用できるものである．

認知心理学における作業記憶に関する研究では，作業記憶の容量に上限があるとされ，一般に 7 個前後であるとされている．ただし記憶する対象によっては 4 個や 5 個程度であることが知られている [9] [10]．

Sarma らは，画像中の物体の個数を数えるタスクにおいて，タスクを分割する際に部分タスクに含まれうる物体の個数を考慮することで，最終的に得られるタスク結果の品質を向上させる手法を提案した [11]．彼らは部分タスク中に含まれる物体の数が 20 から 25 を超えると誤りりつが増加する傾向があると報告した．このように，部分タスクの適切な粒度は対象のタスクやデータの性質によって異なるものの，人間の特性を踏まえたタスク分割はタスク結果の品質管理の観点からも有効な手段と言える．

3 提案手法

ワーキングメモリに基づいてセマンティックセグメンテーションのタスクを分割する手法を提案する．

提案手法ではセマンティックセグメンテーションのタスクをアノテーションを行うカテゴリを単位として複数のサブタスクに分割する．分割することで，複数のクラウドワーカ等に対して並列でタスクを割り当て，それらを統合することで最終的なアノテーションを得ることができる．ここで問題となるのが，1 名のワーカに割り当てるカテゴリの個数の決定方法である．1 名のワーカに割り当てるカテゴリ数が多ければ，それぞれのタスクの作業負担が大きくなるが，一方でカテゴリ数を少なくするほど，各ワーカがタスクの内容を把握するための時間が増加し，さらにアノテーション対象のカテゴリとその周囲のカテゴリとの整合性を取るのが難しくなる．

そこで，認知心理学分野における作業記憶の研究に基づいて各ワーカに割り当てるカテゴリ数を決定する．作業記憶に関する研究では，作業記憶として保持することができる数字や単語などの種類は 7 個前後であると知られている．作業記憶の容量を考慮して，いくつかのカテゴリに対するアノテーションをワーカに割り当てることにより，タスクの過剰な細分化を防ぐことと，割り当てるカテゴリ数が多すぎることによる作業負荷の問題を解決することを目指す．

4 評　　価

ワーキングメモリに基づくタスク分割の有効性を明らかにするために，実験を行った．実験では，1 名のワーカに割り当てるカテゴリ数をグループごとに変化させ，各グループにおけるタスク結果品質と総作業時間を比較した．その結果，カテゴリごとに別のタスクに分割したり，全てのカテゴリで 1 つのタスクとするよりも提案手法に基づくタスク分割が作業時間および作業品質の観点で効率的であることを確認した．

4.1 実験設定

4.1.1 データセット

実験には，セマンティックセグメンテーションのデータセットである ADE20K に含まれる画像から 100 枚の画像を使用した．使用する画像は，次に示す条件に該当する画像の中からランダム抽出により決定した．

- 画像の正解ラベルに少なくとも 1 つの車および道クラスを含む．
- 画像のサイズが 2560×1920 ピクセルである．

4.1.2 アノテーション対象のカテゴリ

実験では，次に示す 20 種類のカテゴリのアノテーションを行った : building, window, tree, signboard, door, person, car, windshield, mirror, sidewalk, handle, road, headlight, wheel, rim, license plate, sky, streetlight, taillight, and plant．これらのカテゴリは，選択した画像に含まれる正解ラベルのうち，出現頻度の高い上位 20 件である．

4.1.3 タスクのインタフェース

実験では，Amazon Web Service が公開しているアノテーションインタフェース作成ツールである Crowd HTML を用いて，セマンティックセグメンテーションのためのアノテーションのインタフェースを作成した．図 2 に実際の作業画面の例を示す．図の例では 5 つのカテゴリが画像の右側に表示されている．実験ではこの部分に表示させるカテゴリの個数を変化させる．

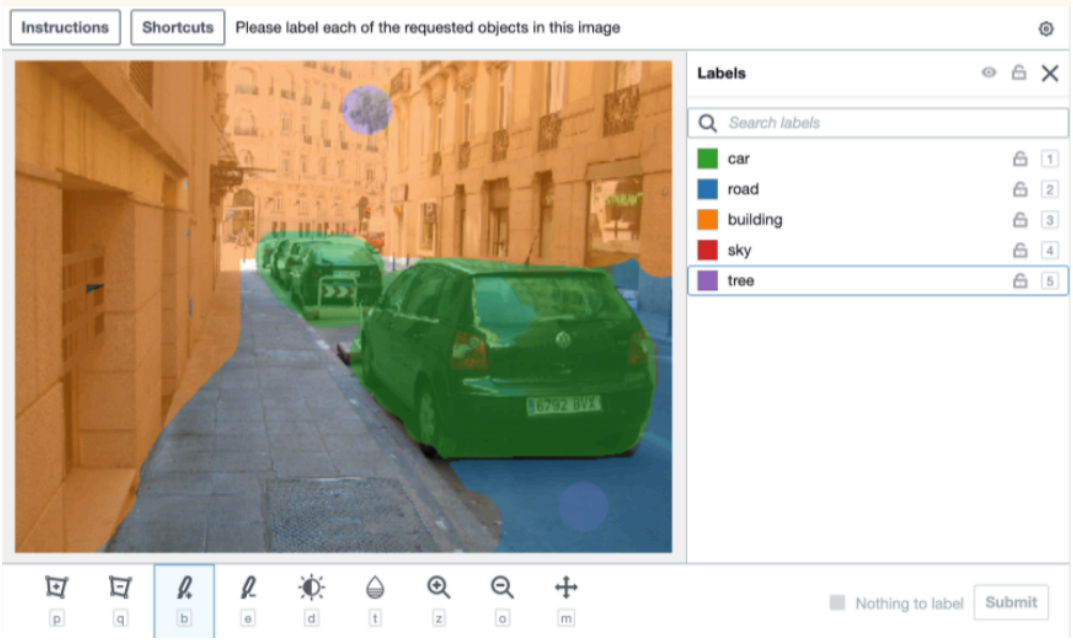

図 2　実験で用いるタスクの例．実験では 1 名のワーカに割り当てるカテゴリ数を変化させる．

4.1.4　Group conditions

実験では，1 枚の画像へのアノテーションの作業の分割方法が異なる 5 つのグループを比較する．表 1 に各グループの詳細を示す．

実験に参加するワーカは，最大で 5 タスクの作業に参加することができる．しかしながら，Amazon Mechanical Turk ではワーカが参加できるタスク数の上限を設けるための仕組みが存在しない．そこで，本実験を実施するにあたり，ブラウザ等に実装されている Web Storage API である Local Storage を用いて，ワーカが本実験に参加した回数を管理した．ワーカが上限を超えて作業を行おうとした場合，作業回数の上限を設けている説明とともに作業画面を無効化する．実験結果の説明で述べるように，この方法によるタスク参加回数の管理方法は厳密ではなく，いくつかのワーカが上限である 5 タスク以上の作業を行なった．

各グループのワーカは，Amazon Mechanical Turk の Qualification Type 機能により管理されており，あるグループのタスクに参加したワーカが別のグループに参加することはできない．

各ワーカは作業が完了した時点で事前に提示された金額の報酬を受け取る．実験では一枚の画像へのアノテーションのコストを$1 とし，各ワーカが受け取る報酬はタスクの分割数に応じて均等に配分される．詳細は表 2 に示す．

4.1.5　評 価 指 標

実験では，タスク結果品質の評価指標として mIoU (Mean Intersection over Union) と pixel-level accuracy の平均値を用いる (式 1)．ただし，y_i は画像データの集合 I の画像 i に対応するピクセル単位でのラベルの集合であり，$y_{i,t}$ は正解ラベルの集合，$y_{i,c}$ はクラウドワーカから得られたアノテーション結果の統合結果の集合とする．関数 $mIoU$ は mIoU の計算結果を返し，関数 $pixelAcc$ は pixel-level accuracy を返すものとする．

$$score(y_{i,t}, y_{i,c}) = \frac{mIoU(y_{i,t}, y_{i,c}) + pixelAcc(y_{i,t}, y_{i,c})}{2} \quad (1)$$

さらに，1 枚の画像へのアノテーションの総作業時間を評価する．

4.1.6　タスク結果の統合

各画像への最終的なアノテーションを得るために，各ワーカから得られたアノテーションを統合する必要がある．実験では，ランダムな順番で各ワーカのアノテーションを重ね合わせて，最終的なアノテーションを作成した．

各ワーカから得られたアノテーションの統合方法は興味深い研究課題の 1 つである．経験則などに基づいてカテゴリごとの平均的な面積を考慮しながら重ね合わせる順番を決定するなどの工夫により最終的なアノテーションの品質を改善できる可能性がある．

5　実 験 結 果

表 3 に各グループごとの参加者と参加したタスク数の内訳を示す．

実験では，各ワーカは最大で 5 タスクに参加できるようにタスクインタフェースに制限を設けた．ただし適用した方法は外部の Web システム等に依存せずに実現可能な簡素な方法であるため，一部のワーカは 5 タスクを超えて作業を行なった．5 タスクを超えて作業することができた理由として挙げられるのは，複数の Web ブラウザや端末を使用していることや，複数の AMT アカウントを所有していることである．

5.1　タスク結果品質

図 3 に各グループごとのタスク結果品質の平均値を示す．実

表 1　Group condition.

Condition (#categories per worker)	#images	#tasks per image	#total tasks	#tasks per worker	#expected workers
1	100	20	2000	5	400
4	100	5	500	5	100
6, 7	100	3	300	5	60
10	100	2	200	5	40
20	100	1	100	5	20

表 2　The reward for a task in each condition.

Condition	Reward per task
1	$0.05
4	$0.2
7	$0.34
10	$0.5
20	$1

表 3　Breakdown of participated workers for each condition

		Assigned tasks per worker		
Condition	#actual workers	Ave.	Min	Max
1	502	3.98	1	87
4	171	2.94	1	39
7	141	2.12	1	10
10	101	1.98	1	10
20	63	1.58	1	6

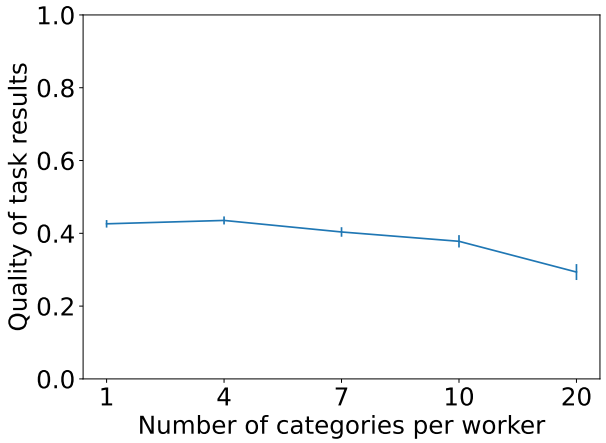

図 3　各グループにおけるタスク結果品質の関係.

験結果から，割り当てるカテゴリ数が 20 に近づくほど，タスク結果品質が低くなる傾向が見られ，また，カテゴリ数が 20 に近づくほどエラーバーで示す標準誤差が微増する傾向がある．

割り当てるカテゴリ数の違いによってタスク結果品質に差があるかを明らかにするために，独立変数をカテゴリ数，従続変数をタスク結果品質として一要因の分散分析を行なった．その結果，カテゴリ数の主効果が認められた ($F(4, 495)=13.871$), $p < .001$) ．Tukey's HSD の方法で多重比較を行ったところ，割り当て数が 20 のグループのタスク結果品質が他の全てのグループよりも低いことが明らかになった ($p < .001$).

このような結果が得られる理由として次の 2 つが考えられる：（1）各ワーカに割り当てるカテゴリ数が増加すると，あるワーカの作業が最終的なタスク結果品質に影響する割合が高くなる．各ワーカに割り当てるカテゴリ数が 1 の場合，仮にあるワーカの作業結果が不正確であっても，最終的なタスク結果品質への影響は少ない．一方で，カテゴリ数が 20 の条件では，1 名のワーカの作業結果そのものが最終的なタスク結果品質となる．ただし，セマンティックセグメンテーションのアノテーションに慣れているワーカが作業する場合，割り当てるカテゴリ数が多くても適当に作業を行うことができる可能性がある．これがカテゴリ数が 20 の条件の標準誤差が高い理由の 1 つと考えられる.（2）各ワーカに割り当てられるカテゴリ数が増えると，各ワーカが取り組むタスクはより複雑で時間を要するものになる．例えば，カテゴリ数が 20 の条件では提示された画像に対して 20 種類のカテゴリで適切な箇所り塗り分ける必要がある．そのため，途中で集中力が途切れてしまうなどの理由により画像全体が適切にアノテーションできなかった可能性がある．

5.2　総作業時間

図 4 に各グループごとの総作業時間（秒）を示す．カテゴリ数が 1 のグループは他のグループより総作業時間が顕著に長く，カテゴリ数が増えるにつれ総作業時間が減少する傾向がある．

グループごとの総作業時間に差があるかを明らかにするために，独立変数をグループ，従属変数を総作業時間（秒）として一要因の分散分析を行った．その結果，カテゴリ数の効果は有意であった ($F(4, 495)=192.364$), $p < .001$)．Tukey's HSD の方法で多重比較を行ったところ，カテゴリ数が 1 のグループの作業時間は他のグループよりも長かった ($p < .001$)．カテゴリ数が 4 のグループの作業時間はカテゴリ数が 7, 10, 20 のグループよりも作業時間が長かった ($p < .001$, $p < .01$, $p < .001$)．また，カテゴリ数が 10 のグループはカテゴリ数が 20 のグループより作業時間が長かった ($p < .001$).

これらの結果から，タスクの分割数が増えると，各ワーカが作業内容を把握するための時間などが増えるなどの理由から，総作業時間が増えたと考えられる．

5.3　タスク結果品質と作業時間の関係

図 5 に各グループのタスク結果品質と作業時間の関係を示す．

各ワーカに割り当てるタスク数を減らすことは，全体的なタスク結果の品質向上に有効である．ただし，総作業時間が増大する．タスク結果品質と作業時間のトレードオフと考慮すると，ワーカに割り当てるカテゴリ数が 4-7 になるようにタスクを分

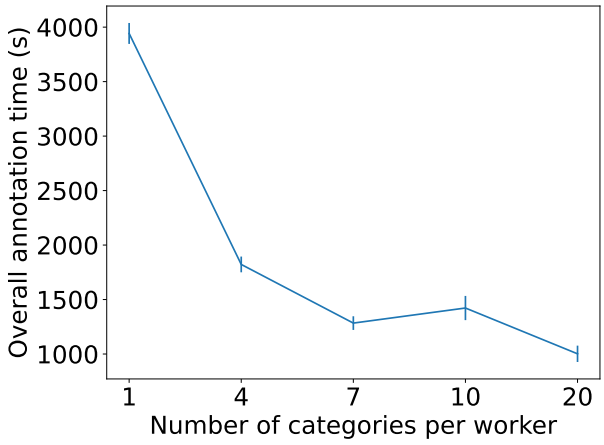

図 4　各グループにおける作業時間の関係.

割することが最も効率的であるといえる.

適切な分割数はアノテーション対象の画像や扱うカテゴリの性質などの影響を受ける可能性があるが，これらを踏まえた分割の最適化は今後の課題の 1 つである.

5.4　アノテーションの失敗例の分析

実験で得られたアノテーションの結果の例を図 6 に示す.

関連研究で挙げたように，アノテータの作業量を削減するなどの目的で，アノテータが全てのピクセルにアノテーションを行う代わりに，点または線による入力を機械学習モデルなどで周囲のピクセルに展開するといった手法が提案されている.

実験結果の一部には，このような手法で求められるアノテーションに類似した作業結果が含まれていた（図 7）. このような作業を行なったワーカは，単に作業を最後まで完了しなかったか，関連研究に類似した実験にも参加した可能性がある.

全てのピクセルへのアノテーションが必要なタスクを依頼する場合は，その旨を明示的にワーカに伝えるといった工夫がタスク結果の品質向上につながる可能性がある.

6　ま　と　め

本研究ではセマンティックセグメンテーションを効率化するために，ワーキングメモリを考慮したタスク分割手法を提案した. ワーキングメモリに基づくタスク分割戦略がカテゴリごとに別のワーカに割り当てたり，全ての作業を 1 名のワーカに割り当てるよりも作業時間やコストの面で有効であった. 今後の課題としてワーカが効率的に作業できるようなカテゴリの選択手法の検討や，複数のワーカから得られたアノテーション結果の統合方法の検討が挙げられる.

文　献

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 39, No. 12, pp. 2481–2495, 2017.

[5] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017.

[6] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, 2016.

[7] Hubert Lin, Paul Upchurch, and Kavita Bala. Block annotation: Better image annotation with sub-image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5290–5300, 2019.

[8] Longlong Jing, Yucheng Chen, and Yingli Tian. Coarse-to-fine semantic segmentation from image-level labels, 2018.

[9] Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, Vol. 24, No. 1, pp. 87–114, 2001.

[10] Nelson Cowan. Working memory underpins cognitive development, learning, and education. *Educational psychology review*, Vol. 26, No. 2, pp. 197–223, 2014.

[11] Akash Sarma, Ayush Jain, Arnab Nandi, Aditya Parameswaran, and Jennifer Widom. Surpassing humans and computers with jellybean: Crowd-vision-hybrid counting algorithms. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 3, pp. 178–187, 2015.

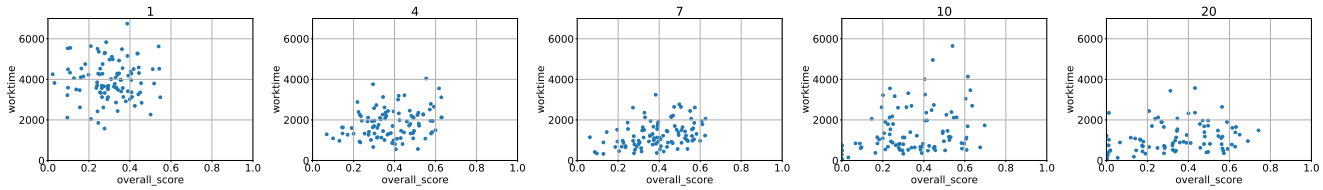

図 5　各グループにおける作業時間とタスク結果品質の関係.

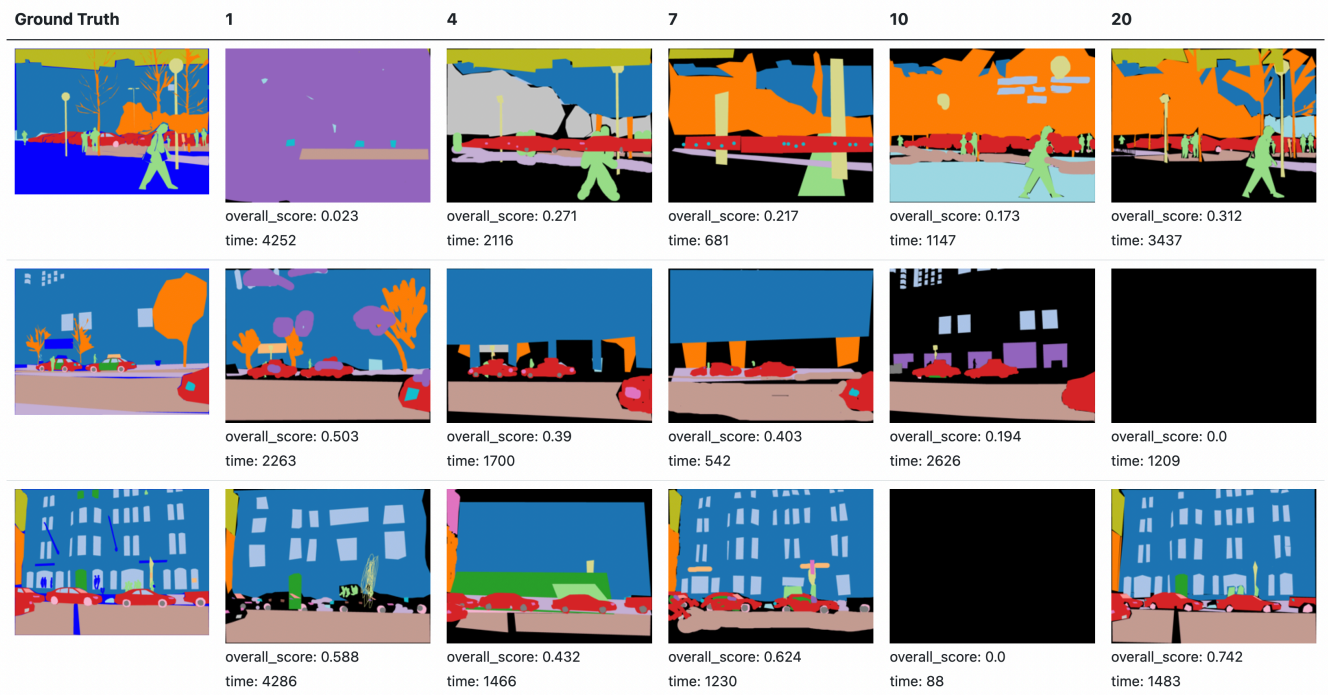

図 6　実験結果の一部.

図 7　依頼者の意図しないアノテーションが付与された例.