Balanced Hierarchical Contrastive Learning with Decoupled Queries for Fine-grained Object Detection in Remote Sensing Images

Supplementary Material

6. Hierarchical Annotations of Datasets

In our experiments, we utilize three fine-grained detection datasets with hierarchical annotations: ShipRSImageNet [34], FAIR1M-v1.0, and FAIR1M-v2.0[23]. Details of the fine-grained categories and their corresponding class hierarchies for each dataset are provided below.

ShipRSImageNet. It contains 41 fine-grained ship categories, including Submarine (SM), Enterprise (EP), Nimitz (NM), Midway (MW), Atago (AT), Arleigh Burke (AB), Hatsuyuki (HS), Hyuga (HG), Asagiri (AS), Ticonderoga (TC), Perry (PR), Patrol (PT), YuTing (YT), YuDeng (YDE), YuDao (YDA), YuZhao (YZ), Austin (AU), Osumi(OS), Wasp (WA), LSD 41 (LSD), LHA, Commander(CM), Medical Ship (MS), Test Ship (TE), Training Ship (TR), AOE, Masyuu AS (MAS), Sanantonio AS (SAS), EPF, Container Ship (CS), RoRo (RR), Cargo (CG), Barge (BG), Tugboat (TB), Ferry (FR), Yacht (YC), Sailboat (SB), Fishing Vessel (FV), Oil Tanker (OT), Hovercraft (HC), and Motorboat (MB). These fine-grained categories are organized into a 4-level class hierarchy by adding 8 coarse-grained ancestor categories, as illustrated in Fig. 7.

Specifically, EP, NM, and MW are categorized as subcategories of Aircraft Carrier (AC); AT, AB, HS, HG, and AS are classified under Destroyer (DES); PR is placed under Frigate (FG); YT, YDE, YDA, YZ, AU, OS, WA, LSD, and LHA are grouped as subcategories of Landing Ship (LA); and MS, TE, TR, AOE, MAS, SAS, and EPF are organized as subcategories of Auxiliary Ship (AUS). Then, SM, TC, PT, and CM, along with the previously introduced coarse-grained categories AC, DES, FG, LA, and AUS, are grouped into Warship. Meanwhile, CS, RR, CG, HC, FR, BG, TB, OT, YC, SB, FV, and MB are grouped into Merchant Ships. Finally, the most general category, Ship, is set as the root node, encompassing both Warship and Merchant Ships. Additionally, each coarse-grained category has an associated "Other" category, which covers instances that cannot be further identified due to low image resolution or those that do not belong to any predefined subcategory. In contrast to existing methods that treat these "Other" categories as mutually exclusive fine-grained categories, we reassign these instances to their corresponding coarse-grained categories.

FAIR1M-v1.0/v2.0. Both FAIR1M-v1.0 and FAIR1M-v2.0 contain 34 fine-grained categories, including Boeing 737 (B737), Boeing 777 (B777), Boeing 747 (B747), Boeing 787 (B787), Airbus A321 (A321), Airbus A220 (A220), Airbus A330 (A330), Airbus A350 (A350), COMAC C919 (C919), COMAC ARJ21 (ARJ21), passenger ship (PS), motorboat (MB), fishing boat (FB), tugboat (TB), engineering ship (ES), liquid cargo ship (LCS), dry cargo ship (DCS), warship (WS), small car (SC), bus (BUS), cargo truck (CT), dump truck (DT), van (VAN), trailer (TRI), tractor (TRC), truck tractor (TT), excavator (EX), baseball field (BF), basketball court (BC), football field (FF), tennis court (TC), roundabout (RA), intersection (IS), and bridge (BR). These fine-grained categories are organized into a 2-level class hierarchy, with 5 coarse-grained categories serving as their ancestors, as illustrated in Fig. 6.

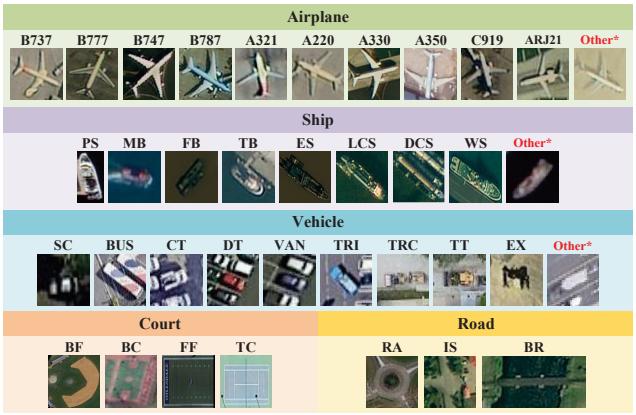

Figure 6. Hierarchical Label Structure of the FAIR1M Dataset. The 34 fine-grained categories are organized into a two-level hierarchy by introducing 5 coarse-grained categories as parent nodes. Additionally, three *Other** categories are included: Other Airplane, Other Ship, and Other Vehicle.

Specifically, B737, B777, B747, B787, A321, A220, A330, A350, C919, and ARJ21 are categorized as subcategories of Airplane; PS, MB, FB, TB, ES, LCS, DCS, and WS are classified under Ship; SC, BUS, CT, DT, VAN, TRI, TRC, TT, and EX are grouped as subcategories of Vehicle; BF, BC, FF, and TC are classified as subcategories of Court; and RA, IS, and BR are organized as subcategories of Road. Additionally, each of Airplane, Ship, and Vehicle includes an "Other" category, and we apply a reassignment procedure similar to that of ShipRSImageNet.

7. Validation of Imbalanced Data Distribution

The experimental results in Section 4.3.2 reveal that, on the FAIR1M-v1.0 dataset, applying the hierarchical contrastive loss (HCL) alone leads to performance degradation, whereas the balanced hierarchical contrastive loss (BHCL)

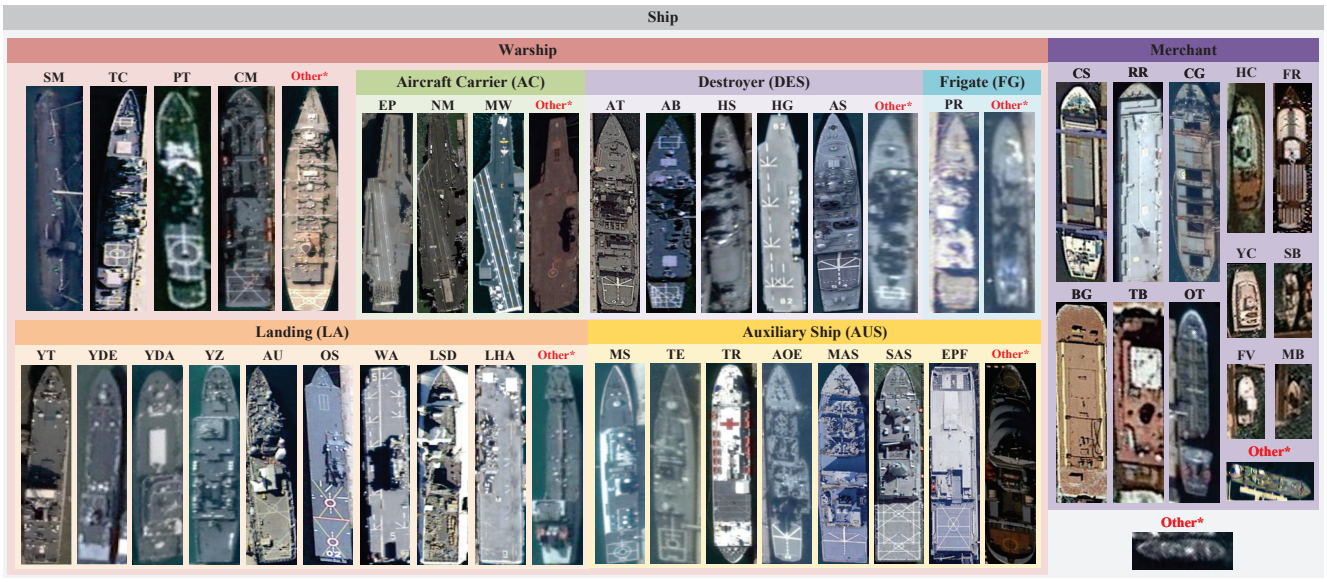

Figure 7. Hierarchical Label Structure of the ShipRSImageNet Dataset. The 41 fine-grained ship categories are organized into a four-level hierarchy by introducing 8 coarse-grained categories as parent nodes. Additionally, eight *Other** categories are included: Other Ship, Other Warship, Other Merchant, Other Aircraft Carrier, Other Destroyer, Other Frigate, Other Landing, and Other Auxiliary Ship.

yields consistent improvements. We attribute this difference to the more severe long-tail distribution in FAIR1M-v1.0 compared to ShipRSImageNet, where HCL becomes dominated by instances from head classes. This hypothesis is supported by an analysis of instance proportions in the training sets of both datasets. As illustrated in Fig. 8, the top three categories in FAIR1M-v1.0 (SC, VAN, and DT) account for over 70% of all instances, significantly outnumbering the remaining categories and confirming the pronounced class imbalance.

8. Resilience to Label Noise

To simulate annotation errors, we conduct an ablation study on ShipRSImageNet by randomly shuffling the parent-child relationships for 10% of the subclasses. The result (64.03 $AP_{50:95}$) exhibits a negligible performance drop of 0.29 compared to the original (64.32 $AP_{50:95}$). This empirically demonstrates that BHCL learns robust visual discriminability rather than overfitting to semantic noise.

9. Optimization Stability under Deep and Imbalanced Settings

Our method scales effectively to the deep 4-level hierarchy of ShipRSImageNet and maintains robustness on the severely long-tailed FAIR1M dataset, successfully overcoming the performance degradation typically observed with HCL. Furthermore, the training AP curves in Fig. 9 confirm that our model converges as smoothly as the baseline, demonstrating consistent optimization stability across both deep-hierarchy and long-tail scenarios.

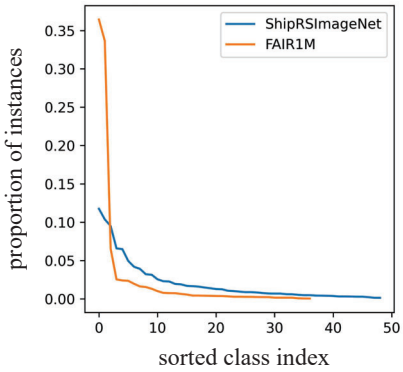

Figure 8. Instance Proportions per Category in the Training Sets of ShipRSImageNet and FAIR1M.

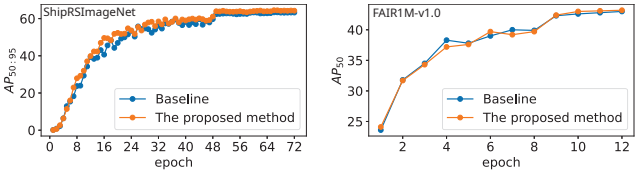

Figure 9. Comparison of training AP curves between our method and the baseline on ShipRSImageNet and FAIR1M-v1.0.

10. Validating the Conflict between Hierarchical Semantic Grouping and Localization

BHCL pulls sibling class queries towards a shared parent prototype, potentially suppressing the spatial variance cru-

Table 4. Comparison of accuracy-efficiency trade-offs between our proposed method and the baseline OrientedFormer (OF).

ShipRSImageNet	**AP**$_{50:95}$	**FLOPs** (G)	**FPS**
Baseline (OF)	63.17	479	21.0
Ours	**64.32** (+1.15)	493 (+2.9%)	20.2 (-0.8)

cial for precise localization. To validate this, adding BHCL without decoupling yields only 63.48 AP$_{50:95}$, whereas our full decoupled method achieves 64.32 AP$_{50:95}$. This 0.84 gap substantiates that decoupling is critical for resolving the conflict between hierarchical aggregation and precise localization.

11. Accuracy-Efficiency Trade-off Analysis

Table 4 compares the computational cost and detection performance of the proposed method against the baseline. Notably, our method achieves a significant gain of 1.15 in AP$_{50:95}$ with only a 2.9% increase in FLOPs and a marginal 0.8 drop in FPS. These results demonstrate a highly favorable accuracy-efficiency trade-off.

12. Visual Comparison of Detection Results

Fig. 10 illustrates the qualitative detection results of OrientedFormer and our proposed method on the ShipRSImageNet and FAIR1M-v2.0 validation sets. Our approach exhibits a lower miss rate for ship targets and achieves superior detection accuracy under misty conditions, demonstrating enhanced overall robustness. However, as highlighted in these visual comparisons, our method still faces limitations in accurately detecting extremely small-scale instances.

13. Fine-Grained Performance Comparisons with State-of-the-Art Methods

In Section 4.4, we present overall performance comparisons between our proposed method and state-of-the-art approaches. Our method outperforms the second-best approach on the ShipRSImageNet, FAIR1M-v1.0, and FAIR1M-v2.0 datasets by $+1.1$ AP$_{50:95}$, $+0.35$ AP$_{50}$, and $+0.49$ AP$_{50}$, respectively. For a more detailed understanding of the performance differences, we provide fine-grained comparison results in Tables 5, 6, and 7. As evident from Tables 6 and 7, our method achieves better performance over the compared methods, particularly in detecting fine-grained categories under Ship and Vehicle.

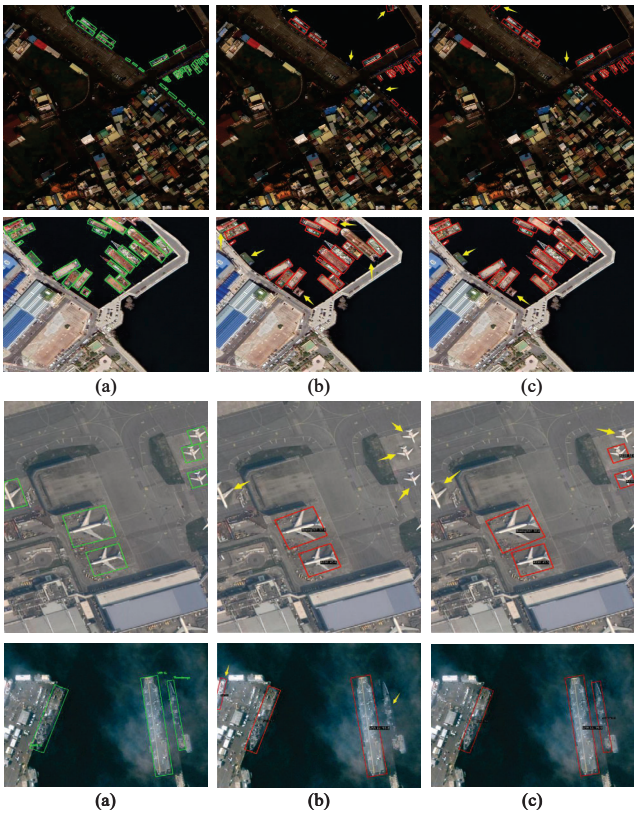

Figure 10. Qualitative comparison of detection results on the ShipRSImageNet and FAIR1M-v2.0 validation sets. Rows 1 and 3 display the results for ShipRSImageNet, while Rows 2 and 4 correspond to FAIR1M-v2.0. The columns represent: (a) Ground Truth, (b) OrientedFormer, and (c) Ours.

Table 5. Comparison of Fine-Grained Detection Results with State-of-the-Art Methods on ShipRSImageNet. All methods employ a ResNet-50 backbone with input images resized to 1024 $\times$ 1024 pixels. Results for the compared methods, except RHINO and OrientedFormer, are adopted from [16].

Method	ReDet	ORCNN	PETDet	PCLDet	LSKNet	SAFPN+APCL	RHINO	OrientedFormer	**Ours**
SM	44.4	**69.0**	46.3	66.9	62.6	68.8	45.5	43.7	43.5
EP	69.6	70.6	64.2	**77.5**	72.4	75.7	69.0	73.5	69.0
NM	67.3	77.4	74.4	74.6	78.6	78.9	65.6	**82.8**	73.3
MW	61.2	**84.8**	74.2	53.3	60.4	71.2	77.6	74.1	81.8
TC	51.8	**83.6**	67.9	82.9	82.8	82.0	62.9	67.6	70.1
AT	33.2	49.3	65.7	57.9	53.1	61.6	64.3	68.8	**70.6**
AB	77.8	87.5	72.7	87.7	85.8	**87.8**	69.3	77.5	75.9
HS	23.2	60.0	66.1	50.9	41.9	56.6	69.1	**70.6**	68.9
HG	84.1	**100.0**	78.8	97.3	97.4	98.2	80.5	81.7	84.9
AS	4.4	18.6	51.3	17.9	18.7	37.8	**66.0**	61.0	60.9
PR	75.3	89.6	70.9	88.6	87.2	**90.2**	68.2	73.8	73.4
PT	21.6	59.8	39.0	41.0	**72.0**	56.2	48.8	54.2	50.7
YT	48.3	61.7	59.7	59.5	**70.3**	66.2	58.6	50.9	65.2
YDE	3.6	**77.1**	62.6	41.8	44.7	59.9	63.4	67.8	68.0
YDA	51.4	54.5	54.9	39.4	**92.4**	64.5	63.9	57.6	62.0
YZ	31.2	67.6	69.4	73.6	73.5	58.2	76.4	**81.8**	74.2
AU	59.2	50.8	72.8	68.8	67.3	64.0	65.4	**76.5**	68.4
OS	81.8	85.7	85.6	**100.0**	97.4	**100.0**	87.1	91.8	92.7
WA	37.3	71.9	83.9	92.7	**97.4**	89.4	90.9	90.9	90.8
LSD	67.7	67.2	66.9	**74.7**	58.9	69.2	63.1	67.8	70.5
LHA	71.8	**89.1**	64.6	88.2	87.3	88.4	69.0	67.8	71.7
CM	64.0	81.6	79.8	**84.6**	77.1	80.4	72.6	81.5	81.8
MS	29.2	51.5	76.6	74.7	62.2	63.3	69.8	72.5	**77.6**
TE	10.4	37.9	54.2	36.6	46.0	41.9	54.0	**71.5**	69.8
TR	30.3	67.4	68.8	**92.9**	73.2	72.3	79.9	79.1	79.5
MAS	13.6	60.9	81.2	74.3	66.9	65.1	85.9	**92.7**	90.9
AOE	7.8	11.4	81.8	27.0	36.2	27.3	84.3	**89.6**	89.5
SAS	47.3	68.7	75.8	72.4	64.7	70.4	75.2	72.0	**76.8**
EPF	**82.7**	81.4	65.2	76.7	67.7	82.1	72.8	78.5	79.6
CS	48.4	50.3	44.4	50.1	47.5	**56.1**	46.7	47.5	47.4
RR	73.4	76.1	62.4	82.1	74.1	**83.1**	69.0	64.9	68.8
CG	58.3	**63.6**	50.5	58.5	55.5	**63.6**	49.6	53.9	58.3
BG	14.6	5.7	17.3	9.9	11.0	13.7	16.6	24.6	**33.3**
TB	46.1	51.0	38.2	55.6	**57.0**	54.3	37.4	37.7	38.0
FR	23.0	22.2	34.1	29.1	19.8	22.7	32.7	38.3	**41.3**
YC	64.6	69.9	56.6	73.5	**73.7**	71.0	53.1	59.7	56.4
SB	15.5	15.8	9.0	16.6	17.1	**18.9**	8.7	16.2	17.5
FV	29.3	**34.3**	18.4	25.0	29.7	31.2	24.2	24.4	28.5
OT	41.0	43.8	46.9	**65.5**	52.0	45.8	40.8	44.3	50.0
HC	64.2	55.9	47.5	**65.8**	50.5	60.4	44.6	47.8	53.5
MB	18.8	16.1	7.1	**21.0**	17.6	20.3	8.6	11.0	11.6
$AP_{50:95}$	45.1	59.5	58.7	61.6	61.0	62.7	59.8	63.2	**64.3**

Table 6. Comparison of Fine-Grained Detection Results with State-of-the-Art Methods on FAIR1M-v1.0. All methods employ a ResNet-50 backbone with input images resized to 1024 $\times$ 1024 pixels. Results for the compared methods, except OrientedFormer, are adopted from [28].

Method	FRCNN	RoITrans	ORCNN	PCLDet	SFRNet	PETDet	DRNet	OrientedFormer	**Ours**
B737	33.94	39.15	35.17	35.96	39.70	**41.05**	39.05	38.46	38.64
B747	84.25	84.72	85.17	**85.50**	84.44	82.23	84.91	83.86	83.57
B777	16.38	14.82	14.57	15.15	17.79	22.04	16.42	**24.22**	21.65
B787	47.61	48.88	47.68	47.97	48.96	**51.21**	49.13	41.30	43.02
C919	14.44	19.49	11.68	16.97	21.18	25.66	**52.50**	17.12	23.36
A220	47.40	50.31	39.05	45.64	48.38	51.83	**68.26**	48.09	47.06
A321	68.82	70.16	39.05	69.14	**71.25**	69.53	66.98	66.62	66.07
A330	**72.71**	70.34	68.60	71.54	72.06	70.86	39.05	65.49	67.39
A350	**76.53**	72.19	70.21	71.50	74.16	73.70	70.27	74.08	70.09
ARJ21	26.59	33.72	25.32	38.68	31.47	36.90	**45.45**	35.32	34.74
PS	11.03	12.62	13.77	17.43	**17.44**	13.29	16.74	15.86	14.19
MB	51.22	55.98	60.42	56.85	60.61	61.94	58.30	63.95	**66.40**
FB	6.41	6.12	9.10	7.88	8.50	8.90	8.08	9.74	**10.11**
TB	34.19	35.31	36.83	36.70	34.87	37.88	32.52	36.95	**39.04**
ES	9.41	9.27	11.32	10.97	**12.55**	10.93	11.97	11.40	12.21
LCS	15.17	15.95	21.86	19.91	19.71	22.05	22.07	22.10	**22.87**
DCS	32.26	34.15	38.22	**40.05**	38.25	36.82	37.74	39.01	39.41
WS	11.27	15.29	22.67	23.80	22.00	23.98	24.79	**28.62**	28.40
SC	54.56	57.55	57.62	56.78	57.73	68.81	58.39	**70.62**	**70.62**
BUS	22.94	26.43	24.40	35.10	32.37	18.33	35.51	36.79	**38.56**
CT	37.74	39.38	40.84	42.19	41.01	42.17	42.90	45.00	**46.26**
DT	41.69	44.95	45.20	45.83	46.69	47.15	47.23	49.47	**50.40**
VAN	48.23	53.69	54.01	52.01	54.08	65.48	53.59	70.86	**70.92**
TRI	12.46	10.95	15.46	**16.34**	15.75	11.93	15.80	11.54	13.16
TRC	2.44	2.13	2.37	6.43	**7.09**	2.09	4.26	1.99	4.85
EX	11.35	10.99	13.55	**18.08**	15.97	8.51	17.42	16.28	17.28
TT	0.32	0.60	0.24	0.61	0.37	0.41	1.05	0.60	**1.29**
BC	45.18	46.93	48.18	48.45	49.43	42.64	50.43	**50.86**	47.56
TC	77.75	79.29	78.45	78.13	79.17	78.60	**82.25**	79.81	79.04
FF	52.05	56.72	60.79	62.55	59.90	**63.51**	60.53	54.97	55.97
BF	87.19	87.21	**88.43**	88.23	87.90	87.40	86.96	85.59	86.61
IS	58.71	58.21	57.90	59.01	59.48	57.48	58.51	**60.42**	59.52
RA	19.38	21.98	17.57	20.21	22.05	20.82	**24.29**	17.97	16.58
BR	20.76	23.31	28.63	31.67	**32.76**	22.57	26.58	29.61	29.68
AP_{50}	36.83	38.49	38.85	40.39	40.74	40.55	40.87	41.31	**41.66**

Table 7. Comparison of Fine-Grained Detection Results with State-of-the-Art Methods on FAIR1M-v2.0. All methods employ a ResNet-50 backbone with input images resized to 1024 $\times$ 1024 pixels. Results for the compared methods are adopted from [28].

Method	FRCNN	RoITrans	GVertex	ORCNN	PCLDet	SFRNet	PETDet	DRNet	**Ours**
B737	42.84	47.76	42.95	44.24	42.77	45.29	**48.96**	43.28	45.04
B747	93.28	94.13	92.98	93.27	93.44	92.97	**94.57**	93.68	93.26
B777	35.63	38.62	37.59	36.81	36.96	40.87	**43.24**	39.87	38.72
B787	62.19	62.21	57.19	59.27	61.60	61.21	**64.50**	63.72	62.13
C919	3.58	12.18	6.31	6.35	12.79	15.96	**26.89**	16.49	11.91
A220	53.51	54.96	53.02	55.29	52.53	54.17	**57.45**	54.20	57.14
A321	68.28	69.07	67.11	70.25	69.76	69.36	**74.41**	68.07	70.61
A330	58.65	60.11	57.98	61.05	**64.11**	61.50	62.84	59.82	62.56
A350	63.46	66.17	65.02	68.04	68.18	68.26	68.45	**70.25**	68.76
ARJ21	10.80	15.34	12.35	10.40	13.52	15.68	11.99	**18.92**	17.62
PS	10.80	13.51	12.75	14.12	14.93	14.45	14.13	**17.00**	14.15
MB	51.82	57.76	56.51	61.61	58.04	61.47	62.88	62.03	**65.56**
FB	12.65	20.89	15.99	26.14	25.12	26.58	24.61	**29.38**	26.22
TB	28.35	30.78	27.63	30.55	30.34	30.56	31.59	**33.01**	32.45
ES	11.45	15.12	13.23	15.97	17.03	17.89	16.53	**18.39**	17.75
LCS	32.95	45.49	40.72	49.45	50.76	48.74	49.88	50.85	**53.01**
DCS	35.34	49.50	42.42	51.95	52.11	50.79	49.46	52.11	**54.26**
WS	13.88	29.06	20.95	33.01	30.22	31.98	28.92	36.78	**38.14**
SC	52.23	58.68	46.62	56.41	56.94	56.77	70.17	58.57	**71.91**
BUS	24.12	30.25	26.47	31.81	34.10	32.58	31.31	**44.71**	35.58
CT	45.51	50.35	44.21	51.29	49.85	50.93	52.91	53.10	**55.26**
DT	44.54	48.80	41.74	48.74	49.71	49.00	51.14	50.36	**54.35**
VAN	47.00	52.62	40.20	52.87	51.24	52.77	68.30	52.87	**72.12**
TRI	8.63	13.18	12.95	15.89	15.81	**17.56**	12.94	16.54	16.45
TRC	1.12	2.73	0.85	1.55	2.46	2.47	0.47	1.83	**7.48**
EX	8.43	13.79	12.12	15.96	17.52	16.20	13.34	**20.85**	17.70
TT	5.93	18.57	6.24	8.06	7.60	29.35	9.96	27.34	**36.33**
BC	55.31	54.71	55.08	59.77	60.65	60.97	53.89	**62.89**	56.92
TC	85.76	86.40	86.70	88.11	87.45	89.35	87.76	**90.47**	88.05
FF	60.48	63.02	58.29	63.70	64.83	61.49	**66.38**	66.30	58.82
BF	91.11	90.93	89.96	91.63	**92.02**	90.56	91.29	91.01	90.45
IS	63.04	65.17	64.34	64.37	**65.85**	65.11	64.67	65.62	64.69
RA	29.98	33.78	32.30	31.52	30.72	**35.13**	33.06	34.13	28.00
BR	25.90	34.07	30.62	36.47	**37.44**	35.05	28.57	34.87	32.64
AP_{50}	39.37	44.11	40.33	44.29	44.66	45.68	46.10	47.04	**47.53**