Category-Aware Aircraft Landmark Detection

Yi Li, Yi Chang, *Member, IEEE,* Yuntong Ye, Xu Zou, Sheng Zhong, and Luxin Yan*, *Member, IEEE*

***Abstract*—Aircraft landmark detection (ALD) aims at detecting the keypoints of aircraft, which can serve as an important role for subsequent applications such as fine-grained aircraft recognition. In ALD, the physical size discrepancy between different kinds of aircraft may lead to inconsistent landmark structure, which significantly harms landmark detection results. In this paper, we take advantage of the category prior to alleviate the size discrepancy in ALD. The proposed category-aware landmark detection network (CALDN) possesses two streams: a classification stream for size categorization and a localization stream for landmark detection. Instance-level size category information captured by classification stream serves as the guidance in the localization stream for robust landmark detection. Moreover, a category attention module (CAM) is proposed for better utilizing category information to guide ALD. Benefitting from the adaptive attention mechanism, CAM can automatically highlight category-specific features for ulteriorly reducing the influence of size discrepancy. Furthermore, to advance ALD research, we contribute the first perspective-variant aircraft landmark dataset. Solid experiments demonstrate the superiority of our method.**

***Index Terms*—Aircraft, landmark detection, category information, convolutional neural networks**

I. Introduction

AIRCRAFT landmark detection (ALD) refers to the task of detecting a set of pre-defined keypoints of aircraft in a given image, which is of great use to numerous subsequent applications such as fine-grained image recognition [1], [2], part-based object recognition [3], and 3D reconstruction [4].

In the past years, only a few works [1], [5], [6] have been proposed to solve ALD problem. Zhao *et al.* [1] predicted aircraft landmark locations directly from the given image through a regression network. However, these works mainly focused on landmark detection only in aerial images, which possess less view variation or self-occlusion due to the single orthographic view. In this paper, we mainly focus on the problem of aircraft landmark detection in variant perspective.

Besides, many works have been proposed to solve keypoints locating problems like human pose estimation [8]–[13] and facial landmark detection [14]–[20]. Most methods aimed at learning a robust feature representation [9], [11]–[13] or establishing geometric relation between joints or landmarks [19], [20]. However, human bodies or faces, which share the same category as they all possess similar components and structures, are different from objects in ALD with significant intra-class size discrepancy between different kinds of aircraft.

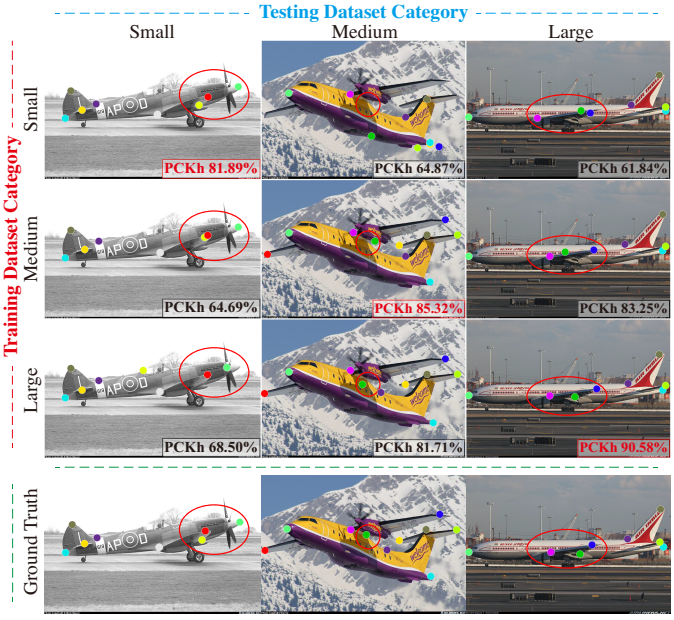


Fig. 1. Analysis of the influence of the fine-grained aircraft category. The aircraft are classified into *large*, *medium* and *small* according to the physical size. In general, aircraft belong to the same size category usually possess structure consistency [7]. Each row and column represent the training and testing data category respectively. Results trained with relevant category of data always perform better and possess less ambiguous prediction than other results, denoting the existance of intra-class size discrepancy.

To illustrate the influence of the aircraft category, we train three different models for aircraft with *large*, *medium* and *small* size categories respectively. The confusing matrix results are shown in Fig. 1. We can observe that introducing the relevant categories for training and testing always leads to the best performance. Furthermore, the closer the training and testing data size categories are, the better performance can be obtained. That is to say, there does exist the size discrepancy between aircraft with different categories. It motivates us to take size category discrepancy among the aircraft into consideration for fine-grained feature representation, so as to facilitate robust aircraft landmark detection.

In this work, we propose a novel category-aware landmark detection network (CALDN) possessing two streams: a classification stream for aircraft size categorization and a localization stream for landmark detection. The category understanding sub-network in the classification stream receives the entire image as input and learns to extract category-specific features. Then the extracted features are merged with landmark features in the localization stream via a specially designed category attention module (CAM), which can auto-

This work was supported in part by the National Natural Science Foundation of China under Grant 61971460, the National Advanced Research Foundation of China under Grant 61406190102 and the China Postdoctoral Science Foundation, No. 2020M672748. (Corresponding author: Luxin Yan)

Yi Li, Yuntong Ye, Xu Zou, Sheng Zhong, and Luxin Yan are with National Key Laboratory of Science and Technology on Multi-spectral Information Processing, the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: li_yi@hust.edu.cn, yuntongye@hust.edu.cn, zx@zoux.me, zhongsheng@hust.edu.cn, yanluxin@hust.edu.cn).

Yi Chang is with the Pengcheng Lab, Shenzhen, China (e-mail: owuchangyuo@gmail.com).

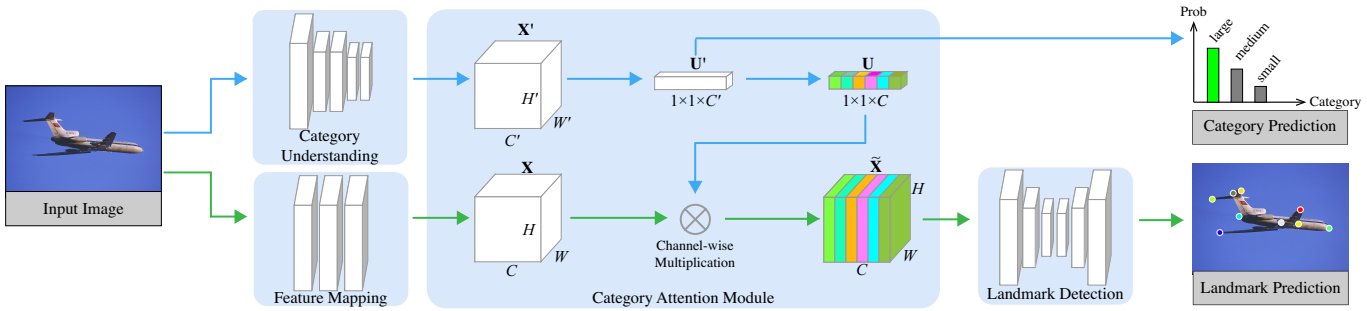


Fig. 2. Overview of the proposed CALDN, which is composed of two streams: a classification stream (blue arrows) for category understanding and a localization stream (green arrows) for landmark detection. Given an input image, category understanding sub-network is used to capture category-specific features, followed by category-aware attention vector generation. Then the learned attention vector selectively highlights the landmark features in localization stream. Finally, the highlighted features are fed into landmark detection sub-network for landmark prediction.

matically highlight category-specific information and generate fine-grained feature representation adaptively. The landmark detection sub-network finally takes the merged features as input to generate accurate landmark prediction. With category information serving as guidance, the proposed CALDN can reduce the feature ambiguous results caused by similar visual appearance and achieve better performance.

In addition, since there are few open datasets specially designed for ALD, we contribute a new perspective-variant aircraft landmark dataset to advance the development of ALD research, which contains 7819 images annotated with 12 landmark locations, bounding boxes and size categories.

Our contributions are summarized in the following aspects. Firstly, we take advantage of category prior for aircraft landmark detection, which endows the network with ability to utilize category information to alleviate size discrepancy and achieve better landmark detection performance eventually. Secondly, we introduce a specially designed category attention module that learns to utilize category information through an adaptive attention mechanism for category-dependent landmark representation. Thirdly, we construct the first perspective-variant aircraft landmark dataset that provides comprehensive landmark annotations for aircraft with different size categories. Extensive experiments demonstrate that our method outperforms the state-of-the-art in both orthographic aerial dataset and perspective-variant aircraft dataset.

II. Category-Aware Landmark Detection

A. Overview of CALDN

Most existing landmark locating methods [9], [12], [13] directly learn the mapping from the input image to landmark location, which mainly exploit the local visual features. However, due to the size discrepancy among different kinds of aircraft, only utilizing local visual appearance information may lead to ambiguous results, especially when there exist high visual similarity parts in the image. To address this problem, we endow the network with ability to understand instance-level size category and enforce landmark predictions to be coherent with the captured category information.

As illustrated in Fig. 2, CALDN is composed of two relevant components, a classification stream for instance-level size category parsing and a localization stream for accurate landmark detection. Specifically, the first component is category understanding sub-network which receives the entire image as input and learns to extract category-specific features $\mathbf{X}'$ automatically. The localization stream is trained to infer landmark locations from the input image, in which feature mapping sub-networks is to map the input image to feature space $\mathbf{X}$. Then the two intermediate features from two streams are merged according to a specially designed category attention module. The merged maps $\widetilde{\mathbf{X}}$ are then fed into landmark detection sub-network to generate landmark prediction coherent with category information extracted by classification stream.

B. Category Attention Module

As size category information is of great importance to ALD, how to utilize category-specific information and merge features are the key issues for accurate landmark detection. There are two straightforward options to merge features from two different streams: channel-wise concatenation [21], and point-wise addition [22]. However, these strategies are unlearnable and cannot adaptively make adjustments according to the input feature [23]. As features from classification stream encode category-specific information and features in localization stream mainly focus on visual appearance, utilizing category-specific features to highlight landmark features dynamically is a reasonable idea. Thus we design a trainable category attention module to adaptively generate category-specific representation for ALD.

Let $\mathbf{X}' \in \mathbb{R}^{H' \times W' \times C'}$ and $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ be the features from two network streams. As shown in Fig. 2, given both $\mathbf{X}'$ and $\mathbf{X}$ as input, CAM first squeezes the category-specific features $\mathbf{X}'$ by global average pooling, followed by a fully connected layer with activation to obtain category attention vector $\mathbf{U} \in \mathbb{R}^{1 \times 1 \times C}$. Then the merged maps are generated by $\widetilde{\mathbf{X}} = \mathbf{X} \bigotimes \mathbf{U}$, where $\bigotimes$ denotes the channel-wise multiplication. Finally, CAM exports the merged maps to generate landmark prediction.

Benefitting from the learnable attention mechanism, CAM can adaptively highlight features through category-specific information and endow algorithm with ability to handle aircraft with category variation in a single network. With category information merged, the network can automatically constrain landmark prediction to be coherent with instance-level guidance. Furthermore, we adopt extra category supervision on the

(a) w/o CAM (b) w/ CAM (c) Ground Truth

Fig. 3. The effectiveness of CAM. Without category information encoded by CAM, algorithm may locate landmarks wrongly at another instance (*e.g.* landmarks on head and left wing are wrongly located on another aircraft). On the contrary, our method can acquire more precise results with the help of CAM which leverages category specific information to constrain landmarks belonging to single aircraft.

top of $\mathbf{U}'$ to explicitly guide the classification stream to distill category-specific features.

To illustrate the effectiveness of CAM, we visualize the results with and without CAM in Fig. 3. Without CAM providing size category information, classical landmark detection methods are more likely to obtain ambiguous results when facing overlap of aircraft in the given image. On the contrary, with category information provided in CAM as constraint, the network is capable of suppressing ambiguous landmark responses on other similar parts. The results strongly validate the effectiveness of CAM.

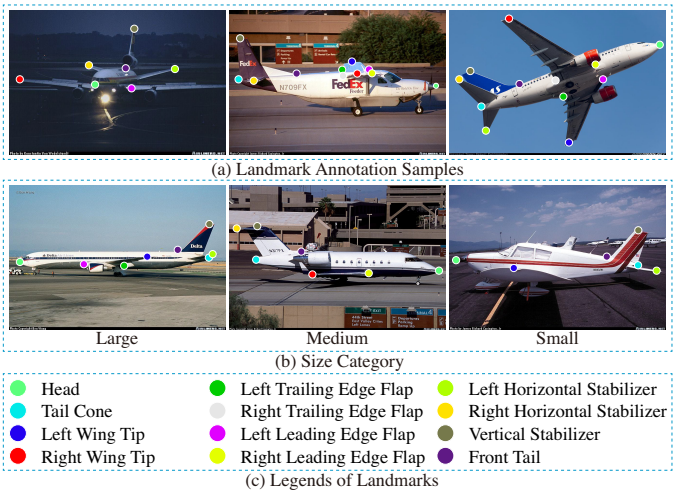
C. Implementation Details

CALDN is end-to-end trainable. We use resnet18 [24] and HRNet-w32 [13] as our category understanding and landmark detection sub-networks respectively. Our feature mapping sub-network consists of three conv blocks, and each conv block possesses a convolution (with 3×3 kernel), a batch normalization layer, and a ReLU activation layer. The $\mathbf{C}'$ and $\mathbf{C}$ in the CAM are set to 3 and 64 respectively. Localization stream and classification stream are supervised by MSE loss ($\mathcal{L}_l$) and cross-entropy loss ($\mathcal{L}_c$) respectively. The total loss $\mathcal{L}_{total}$ can be minimized via $\mathcal{L}_{total} = \omega_1\mathcal{L}_l + \omega_2\mathcal{L}_c$, where ω_1 and ω_2 are balance parameters, which are set to be 1 and 1e-3.

III. Perspective-Variant Aircraft Landmark Dataset

Since few datasets are designed for perspective-variant aircraft landmark detection, in this paper, we construct the first perspective-variant aircraft landmark dataset (PVALD) to benchmark and advance the development of aircraft landmark detection. We select a subset of 7819 images (6246 for training and 1573 for testing) with different size categories and pose variations from FGVC-Aircraft dataset [25] and annotate each aircraft with 12 landmark locations, visibility[1], along with the bounding box of each instance. The definition of the 12 landmarks is shown in Fig. 4(c). We define landmarks mainly on peaks and joints of aircraft parts for a better description of aircraft structure and pose. Sample images and annotations are shown in Fig. 4(a). To provide category information, we divide the images into three subsets according to the physical size of the aircraft (similar to *Aircraft Design Group Classification* [7]), including the subset of *large/medium/small* size. Sample images belong to the different categories are illustrated in Fig. 4(b), which shows that PVALD contains substantial images with large pose and category variations.

[1]Three states of visibility are defined for each landmark, including visible (located inside of the image and visible), invisible (inside of the image but occluded), and outside (located outside of the image).



(a) Landmark Annotation Samples

(b) Size Category

(c) Legends of Landmarks

Fig. 4. Illustration of landmarks and category definition in PVALD. (a) sample images and annotations for different aircraft (only visible points are shown). (b) sample images from different subsets of *large/medium/small* size category. (c) legends of 12 landmarks in PVALD.

IV. Experiments

A. Experiments Setting

Evaluation Metrics. We employ two metrics to evaluate the performance of aircraft landmark detection, PCKh [26] and normalized mean error (NME). PCKh is the percentage of detected landmarks which fall within the neighborhood of the ground truth, while the NME is defined as the mean l_2 normalized distance between predicted landmarks and ground truth. Typically, a higher PCKh score or smaller NME value denotes better landmark detection results.

Datasets. We evaluate and report the results on two datasets. PVALD is our contributed perspective-variant aircraft landmark dataset described in Section III. DOTA-aircraft is a landmark dataset which contains 146 aerial images selected from DOTA [27]. Each aircraft instance is annotated with the bounding box information and the location of 12 landmarks possessing the same definition as PVALD.

Competing Methods Since there are few works specially designed for aircraft landmark detection, we compare CALDN with five state-of-the-art methods in human pose estimation and facial landmark detection, including Hourglass [9], PoseAttention [10], PyraNet [11], SimpleBaseline [12], and HRNet [13]. For a fair comparison, all competing methods are fine-tuned on the corresponding dataset.

Training Details. The whole network is trained on four NVIDIA GeForce GTX 2080Ti GPU with 11GB memory. We use adam [28] as the optimizer with the initial learning rate of 1e-3. We first train category understanding network and landmark detection network separately, and then use the total loss function to fine-tune the whole network. Our source code and dataset will be available on our homepage.

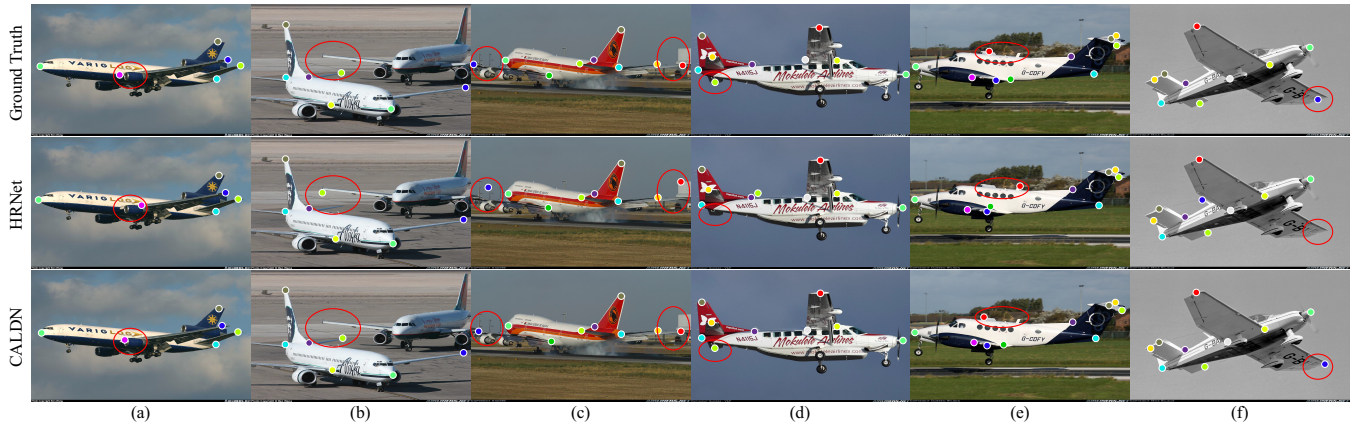


Fig. 5. Visual comparison of aircraft landmark detection by different methods. Top row shows the ground truth of the landmarks. Middle and bottom rows show the results of HRNet [13] and CALDN respectively.

TABLE I
COMPARISONS OF PCKH@0.5 SCORES AND NORMALIZED MEAN ERROR (NME) ON PVALD.

Method	Head & Tail cone	Leading edge flap	Wing tip	Trailing edge flap	Horizontal stabilizer	Vertical stabilizer	PCKh	NME
Hourglass [9]*	96.60	79.55	82.30	72.93	83.13	87.95	86.18	0.4140
PoseAttention [10]	97.27	81.87	77.05	78.54	86.40	89.29	87.16	0.3256
PyraNet [11]	97.15	82.69	77.59	**81.14**	85.06	88.69	87.17	0.3247
SimpleBaseline [12]	97.27	84.25	82.48	76.80	89.32	89.32	88.58	0.3243
HRNet [13]	97.57	85.28	84.13	75.07	90.75	89.53	89.15	0.2952
CALDN	**97.77**	**85.96**	**84.90**	76.42	**91.37**	**89.76**	**89.67**	**0.2905**

TABLE II
COMPARISONS OF PCKH@0.5 SCORES AND NORMALIZED MEAN ERROR (NME) ON DOTA-AIRCRAFT.

Method	PCKh	NME
HRNet [13]	89.15	0.3584
CALDN	93.41	0.3575

B. Evaluation Results on PVALD

Qualitative Evaluation. Figure 5 shows ALD results on PVALD. From the first to the third row, we show the ground truth landmark location, and the results of HRNet [13] and CALDN. With instance-level category information, our method obtains more accurate results and eliminates the wrong predictions caused by similar visual appearance while HRNet may wrongly locate landmarks to other parts of the aircraft. These results strongly support the effectiveness of CALDN.

Quantitative Evaluation. Table I reports the PCKh@0.5 and NME of the competing methods. Note that there are 12 landmarks in our dataset, we merge the scores from the same semantic parts (e.g. left and right leading edge flap are merged into leading edge flap). We can observe that CALDN performs better than the competing methods for most parts, and outperforms the SOTA by 0.52% on PCKh, which further demonstrates the effectiveness of our method.

C. Evaluation Results on DOTA-Aircraft

Qualitative Evaluation. We show the qualitative results of CALDN on DOTA-aircraft [27] in Fig. 6. The first row and second row show aircraft landmark detection results of different size categories. Compared with our PVALD, aircraft in aerial images possess less perspective variation and self-occlusion, which is less challenge for landmark detection.

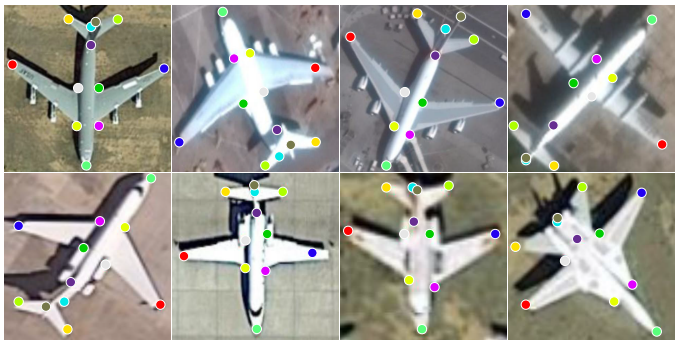


Fig. 6. Results of the CALDN in DOTA-aircraft. The first row represents the results of *large* aircraft, the second row represents *medium* and *small* aircraft.

Quantitative Evaluation. We also present the quantitative results of HRNet [13] and CALDN on DOTA-aircraft in Table II. We can observe that CALDN outperforms the competing method by a large margin, which further supports the effectiveness of our model. And the result of DOTA-aircraft further demonstrates that it is easier to obtain better ALD performance in aerial images than perspective-variant images.

D. Ablation Study

Category Information Fusion and Supervision. We further conduct experiments to investigate the influence of information fusion strategy and category supervision in CALDN. As shown in Table III, the first row denotes the baseline model. The second row represents simply concatenating category information along with the input image. The third and last row

TABLE III
EFFECTIVENESS ANALYSIS OF EACH COMPONENT. CI REPRESENTS THE CATEGORY INFORMATION. CAM AND CS DENOTE THE CATEGORY ATTENTION MODULE AND CATEGORY SUPERVISION.

CI	CAM	CS	PCKh	NME
			89.15	0.2952
✓			89.12	0.2931
✓	✓		89.41	0.2943
✓	✓	✓	**89.67**	**0.2905**

TABLE IV
CLASSIFICATION EFFECTIVENESS AND ORACLE ANALYSIS OF CALDN.

	Classification Accuracy	PCKh	NME
CALDN	94.22	89.67	0.2905
oracle	—	89.73	0.2905

represent CALDN without or with the category supervision respectively. We can observe that category information and CAM does help the network to generate category-specific representation and obtain better results. Benefitting from the adaptive feature selection, result with CAM outperforms that of simply concatenating the category information by a large margin, strongly demonstrating the effectiveness of the attention scheme designed in CAM. Moreover, explicitly utilizing category supervision as constraint also helps the network to distinguish size discrepancy and gains better results for landmark detection.

Oracle Analysis. In Table IV, we compare our method together with the *oracle* that has the access to classification groundtruth information serving as upper bound. With quite accurate guidance provided by category-parsing network, our CALDN result is not far from that of the oracle. Moreover, comparing the upper bound performance represented by the oracles, it is possible to know that accurate category information can indeed improve ALD performance, which further supports the importance of category prior for aircraft landmark detection.

V. CONCLUSION

In this work, we present a novel category-aware network for aircraft landmark detection, which adopts size category information to alleviate the size discrepancy among different kinds of aircraft. We incorporate the category knowledge via CAM to endow network the ability to highlight features for robust landmark detection adaptively. Moreover, to advance the development of ALD, we propose a new perspective-variant aircraft landmark dataset. Solid improvements over SOTA methods demonstrate the effectiveness of CALDN.

REFERENCES

[1] A. Zhao, K. Fu, S. Wang, J. Zuo, Y. Zhang, Y. Hu, and H. Wang, "Aircraft recognition based on landmark detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1413–1417, 2017.

[2] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. ICCV*, 2017, pp. 5209–5217.

[3] K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun, "Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images," *Remote Sens.*, vol. 11, no. 5, pp. 544, 2019.

[4] R. Zhao, R. Wang, and A. M. Martinez, "A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3059–3066, 2017.

[5] J. Zuo, G. Xu, K. Fu, X. Sun, and H. Sun, "Aircraft type recognition based on segmentation with deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 282–286, 2018.

[6] Y. Zhang, H. Sun, J. Zuo, H. Wang, G. Xu, and X. Sun, "Aircraft type recognition in remote sensing images based on feature learning with conditional generative adversarial networks," *Remote Sens.*, vol. 10, no. 7, pp. 1123, 2018.

[7] M. Sadraey, *Aircraft design: A systems engineering approach*, John Wiley & Sons, 2012.

[8] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. ICCV*, 2013, pp. 3487–3494.

[9] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.

[10] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, 2017, pp. 1831–1840.

[11] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. ICCV*, 2017, pp. 1281–1290.

[12] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. ECCV*, 2018, pp. 466–481.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, 2019, pp. 5693–5703.

[14] T. Baltrusaitis, P. Robinson, and L. P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proc. ICCVW*, 2013, pp. 354–361.

[15] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proc. CVPRW*, 2017, pp. 2025–2033.

[16] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," in *Proc. ECCV*, 2018, pp. 339–354.

[17] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *Proc. CVPR*, 2018, pp. 781–790.

[18] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. CVPR*, 2018, pp. 2129–2138.

[19] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao, and L. Lin, "Layout-graph reasoning for fashion landmark detection," in *Proc. CVPR*, 2019, pp. 2937–2945.

[20] X. Zou, S. Zhong, L. Yan, X. Zhao, J. Zhou, and Y. Wu, "Learning robust facial landmark detection via hierarchical structured ensemble," in *Proc. ICCV*, 2019, pp. 141–150.

[21] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[22] R. Wen, K. Fu, H. Sun, X. Sun, and L. Wang, "Image superresolution using densely connected residual networks," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1565–1569, 2018.

[23] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6038–6051, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[25] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[26] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, 2014, pp. 3686–3693.

[27] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proc. CVPR*, 2018, pp. 3974–3983.

[28] P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.