DGV-Suite: A Benchmark Suite of Task-Specific Annotations for Dairy Goat Vision

Yifan Wei
weiyifan@nwafu.edu.cn
Northwest A&F University
Yang ling, Shaanxi, China

Jinglei Tang*
tangjinglei@nwsuaf.edu.cn
Northwest A&F University
Yang ling, Shaanxi, China

Gaoge Han
hangaoge@gmail.com
MBZUAI
Abu Dhabi, UAE

Hao Rong
ronghao@nwafu.edu.cn
Northwest A&F University
Yang ling, Shaanxi, China

Xianglong Pei
peixianglong@nwafu.edu.cn
Northwest A&F University
Yang ling, Shaanxi, China

Ruizi Han
25B951126@stu.hit.edu.cn
Harbin Institute of Technology
Shen Zhen, Guangdong, China

Abstract

Dairy goats have attracted increasing attention recently due to their nutritional and economic importance, as well as their growing role in global livestock production. Meanwhile, computer vision and intelligent multimedia analysis have shown strong potential for improving dairy goat management by enabling automated monitoring, behavior understanding, and fine-grained decision support in practical farming environments. However, existing dairy goat datasets are typically small in scale, task-isolated, and collected under constrained conditions, making them insufficient for intelligent multimedia analysis in realistic farming environments. To address this gap, we present DGV-Suite a large-scale benchmark suite for dairy goat vision, organized as multiple task-specific annotations under a standardized benchmark framework, containing 32,007 images and 1,459 videos collected from authentic farm environments, covering diverse conditions of illumination, occlusion, and inter-animal interaction. The benchmark supports eight representative tasks, including object detection, instance segmentation, semantic segmentation, object tracking, pose estimation, behavior recognition, individual identification, and image generation; it provides a task-specific annotation for diverse dairy goat visual understanding scenarios. We further establish benchmark settings and report baseline results on six supervised tasks using representative models, confirming the practical application challenges and benchmark value of DGV-Suite which provides a valuable resource for advancing future research in precision livestock farming, dairy goat scene understanding, and intelligent multimedia analysis. DGV-Suite is available: https://y-f-wei.github.io/DGV-Suite/

Keywords

Dairy goat vision benchmark, Task-specific annotations, Intelligent multimedia analysis

1 Introduction

Dairy goats are specialized caprine breeds selectively improved for milk production and constitute an important component of diversified livestock and dairy systems worldwide. They are particularly valued for their strong ecological adaptability, efficient use of relatively low-quality feed resources, and capacity to remain productive under environmental and resource constraints. Goat milk has long been integrated into human diets due to its excellent nutritional profile, supporting a wide variety of fresh and processed dairy products. These characteristics make dairy goats not only economically relevant to rural livelihoods and dairy industries but also scientifically important for data-driven livestock research[11, 13, 27].

According to the Food and Agriculture Organization of the United Nations (FAO) world stock statistics from 2000 to 2024 [15], goats exhibit the most pronounced long-term growth among the four major livestock species, compared with the more moderate trends observed for cattle, swine, and sheep. The global goat population exceeded 1.1 billion heads in 2024, representing a 57.6% increase relative to 2000, which further highlights the growing importance of goats in modern livestock production. With the rapid advancement of precision livestock farming (PLF), intelligent multimedia analysis has become a key enabler for automating dairy goat monitoring, health assessment, and individual management[40]. Conventional identification relies on physical markers such as ear tags or RFID chips. Video-based and image-based multimedia sensing provides non-contact continuous monitoring of individuals and enables behavioral analysis[17, 36]. Such approaches can reduce routine handling, lower animal stress, and lessen the burden of manual data collection.[33, 55]. However, compared to cattle and swine, dairy goats remain underrepresented in vision-based and multimedia-oriented research[4]. A major bottleneck is the absence of large-scale, standardized benchmark resources, which hinders the development, evaluation, and fair comparison of robust models for dairy goat visual understanding[31, 32].

Despite growing interest in vision-based and multimedia-enabled livestock analysis, currently available dairy goat datasets remain far from adequate. Existing datasets for dairy goat analysis suffer from several fundamental limitations. First, most available resources are task-isolated, typically designed for a single objective such as object detection[20, 45, 50], pose estimation[29, 39], or individual identification[1, 37], rather than being organized under a common benchmark framework for broader visual understanding in intelligent farming scenarios. This fragmentation limits the development and fair evaluation of models across related tasks and reduces their practical utility in multimedia-enabled livestock management systems. Second, many existing datasets remain relatively small in scale and limited in diversity, often containing only a few thousand images collected under constrained conditions[1, 2, 20, 29, 39, 48]. Such

*Corresponding author

Table 1: Public image and video datasets for livestock farming.

Dataset	Images	Videos	Object Detection	Identification	Pose Estimation	Segmentation	Tracking	Behavior Recognition
FriesianCattle [Andrew et al., 2017]	940	-	✓	✓				
Aerial-livestock [Han et al., 2019]	89	-	✓					
NWAFUCattle [Li et al., 2019a]	2,134	-			✓			
CattleVideo [Qiao et al., 2021]	-	363		✓				
CVB [Zia et al., 2023]	-	502						✓
CattleEyeView [Ong et al., 2023]	-	14	✓		✓	✓	✓	
PigPosture [Riekert et al., 2020]	305	-			✓			
AggressiveBR [Gao et al., 2023]	-	1,106						✓
SheepBreed [Abu Jwade et al., 2019]	1,642	-		✓				
SheepActivity [Kelly et al., 2024]	-	417						✓
DroneGoat [Vayssade et al., 2019]	517	-						✓
RTGoat [Sun et al., 2024]	-	123					✓	
DGV-Suite(our dataset)	32,007	1,459	✓	✓	✓	✓	✓	✓

limitations restrict model robustness and lead to poor generalization across farms, seasons, and breeds. Moreover, dairy goats present distinctive challenges for visual analysis: their high inter-individual similarity makes identity-related tasks particularly difficult, while the lack of long-term annotated video data with persistent individual IDs hinders continuous tracking, behavior understanding, and individualized management.

To address these challenges, we introduce DGV-Suite a large-scale benchmark suite specifically designed for intelligent dairy goat farming and multimedia-based animal understanding. Rather than a fully aligned sample-level multi-task dataset, DGV-Suite is organized as eight task-specific subsets with 32,007 images and 1,459 videos. Compared with existing dairy goat resources, DGV-Suite provides broader and more systematic task coverage under a unified data organization and evaluation framework. Importantly, the dataset was collected in authentic and challenging farm environments, encompassing multiple dairy goat habitats, including barns, exercise yards, and milking parlors, thereby ensuring that downstream models exhibit robustness and high generalization capabilities. In summary, our contributions are as follows:

- **DGV-Suite Task-Structured Benchmark Suite.** We introduce DGV-Suite a large-scale benchmark suite of task-specific annotations for dairy goat vision, covering eight representative tasks in intelligent farming scenarios. DGV-Suite is the first benchmark resource to provide such broad and systematically organized task coverage for dairy goat visual understanding.
- **Task-Specific Annotation Framework.** Rather than formulating all tasks as a fully aligned multi-task dataset, we organize DGV-Suite as eight task-specific annotations under a standardized benchmark framework. Each annotation is constructed with emphasis on precision, definition consistency, temporal continuity, instance discrimination, and real-scene diversity, enabling reliable evaluation across authentic farm environments.
- **Baselines and Insights.** We establish extensive benchmark results across six supervised tasks using representative state-of-the-art architectures. These results validate the complexity and practical value of DGV-Suite while highlighting key challenges, including cross-scene robustness, identity-aware analysis, long-term tracking, and behavior understanding in real farm environments.

DGV-Suite addresses the long-standing lack of standardized benchmark resources for dairy goat-centered visual understanding and intelligent multimedia analysis. We believe it will accelerate research in intelligent animal monitoring and contribute to more efficient and sustainable dairy goat production.

2 Related works

General-purpose vision benchmarks such as ImageNet[12], COCO[30], and AVA[19] have established the foundation for vision recognition, detection, and action analysis. More recent datasets like Animal Kingdom[34] extend to wildlife perception and multi-species behavior understanding.

Recent years have seen increasing adoption of computer vision in precision livestock farming, enabling non-contact monitoring for individual identification, behavior understanding, and welfare assessment. We introduce several easily accessible PLF datasets in Table 1. Cattle-focused datasets have been widely used for re-identification, object detection, and behavior detection, such as the FriesianCattle[2], CattleVideo[37], and CVB[61], yet they lack fine-grained temporal annotations or cross-farm generalization. Pig-oriented datasets, including PigPosture[39], and AggressiveBR[18], provide pose estimation and action behavior detection but are often limited to controlled indoor environments with homogeneous lighting. Despite this progress, several persistent gaps remain. Most livestock datasets are single-task or specific-task, supporting only one perception objective, such as detection, pose estimation, or re-identification, without integrating multiple vision functions[1, 20, 29, 37, 39, 44, 48, 61]. Dataset scales remain modest, typically ranging from a few hundred to thousands of samples[1, 2, 20, 29, 37, 39, 44, 48, 61]. CattleEyeView supports multiple vision tasks, but it only has 14 video sequences[35], insufficient for large-model training or robust cross-domain transfer. Additionally, health-related vision labels[56], such as lameness[53], are rare, restricting their use in welfare monitoring and precision diagnostics. Finally, only a few datasets include temporal continuity[35, 61], a prerequisite for modeling long-term behaviors and detecting subtle anomalies.

Compared to other livestock, dairy goats have lower coverage in public datasets. In recent years, research on dairy goats has gradually increased. RTGoat[44] contains 123 video sequences primarily used for real-time tracking tasks of dairy goats. These datasets are relatively small, collected and annotated for specific tasks only[44, 48, 53], resulting in limited vision diversity and annotation consistency. As a result, they provide insufficient scene

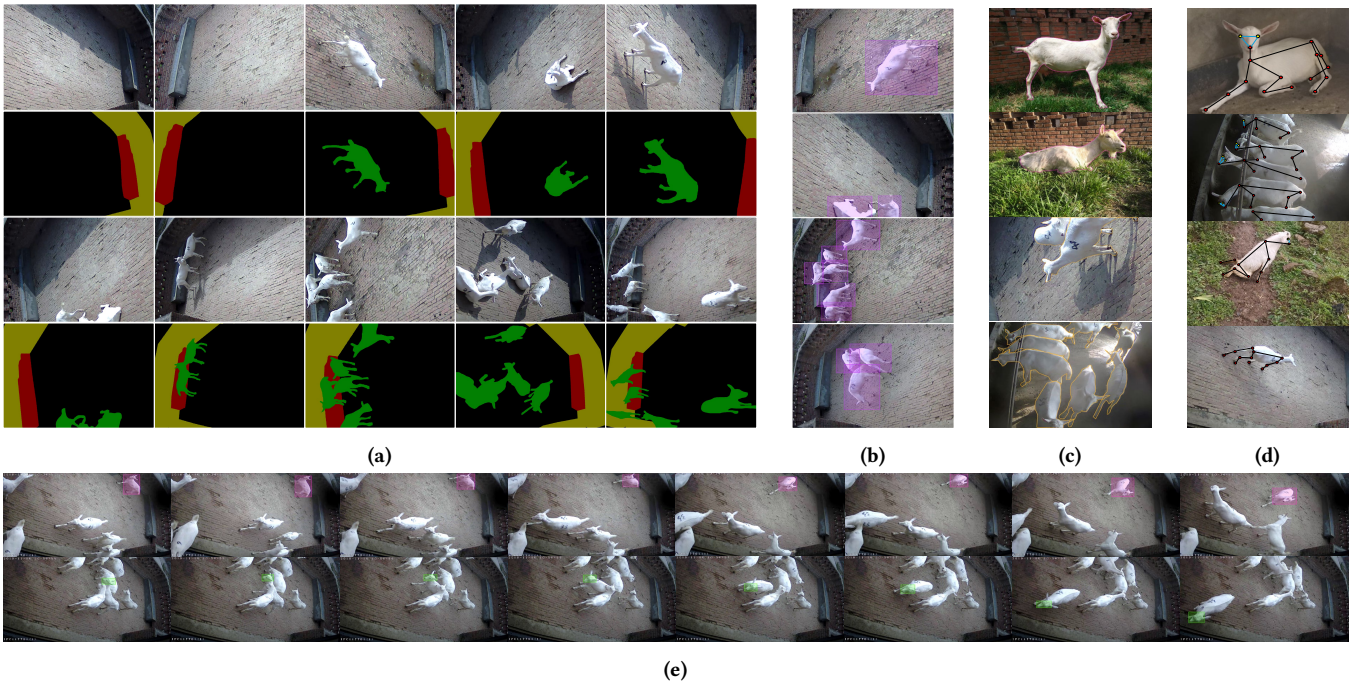

Figure 1: (a) Examples of mask annotations in the semantic segmentation. Green pixels represent dairy goats. (b) Examples of bounding box annotations in object detection. (c) Examples of polygon annotations in the instance segmentation. (d) Examples of keypoint annotations in the pose estimation. (e) Examples of body and head sequences in object tracking.

diversity, annotation consistency, and benchmark coverage for developing robust models or supporting systematic evaluation across broader dairy goat visual understanding scenarios.

3 Dataset construction

3.1 Data acquisition

Farm environment. All data in DGV-Suite was collected at a large-scale dairy goat research and teaching farm in northwestern China. The farm maintains 159 lactating Saanen dairy goats aged from one month to eight years, together with 10 black goats under the same management conditions. The farm consists of a semi-open housing system with both indoor and outdoor areas. The indoor barn is approximately 50 m long, 10m wide, and 5m high, with multiple fenced pens arranged on both sides of a central passage. Each pen is connected to an outdoor enclosure through a small gate with windows above it for natural lighting. The outdoor area is approximately 25m $\times$ 5m and includes feeding troughs for routine activity and feeding. This configuration provides substantial real-world variation in lighting, background, viewpoint, occlusion, and behavioral context.

Acquisition devices and configuration. Data acquisition was performed using a multi-camera configuration deployed in indoor barns and outdoor exercise areas. Continuous videos were recorded using YW7100HR-SC9 high-definition infrared network cameras mounted at elevated indoor positions and beneath outdoor overhangs, providing stable top-down views for long-term monitoring. Additional DS-IPC-B12-I/POE cameras were installed on surrounding support walls at a height of approximately 3m and oriented downward at an angle of about 60° to ensure full coverage of the pens. All cameras recorded at 1080p resolution and 25 fps, and were connected via wired links to a central workstation for synchronized data collection.

During the daily recording period from 09:00 to 18:30, groups of goats moved freely within indoor pens and outdoor enclosures, generating naturally diverse visual scenes. The raw footage was segmented into shorter clips, and clips affected by severe motion blur, defocus, or heavy occlusion were manually removed. To complement the video data, high-resolution still images were captured using a Canon EOS 60D digital camera, with each image centered on an individual goat. Images with blur, low contrast, or strong environmental obstruction were also filtered through manual inspection.

3.2 Annotation protocols

Pose estimation and behavior recognition. For the pose estimation subset, videos were processed with FFmpeg to extract key frames at 5-second intervals, and each frame was annotated with 17 anatomical keypoints in COCO format. The keypoint scheme was designed to be biologically meaningful, consistently defined, and practically annotatable. Annotation quality was controlled with emphasis on keypoint accuracy, pose diversity, and challenging cases such as occlusion, deformation, and viewpoint variation. For the behavior recognition subset, behavior categories were defined

according to the principles of universality, authenticity, and evaluability. Only clips with complete and visually clear action segments were retained, and each segment was manually assigned to a predefined class. To improve the practical value of this subset, we emphasized clear behavior definitions, temporal annotation reliability, behavioral diversity, environmental robustness, and class balance. The detailed definitions of anatomical keypoints and behavior categories are provided in the supplementary material on the DGV-Suite project page.

Detection and tracking. For the object detection subset, video frames were manually annotated with bounding boxes using LabelImg and stored in PASCAL VOC format, with emphasis on accurate localization under occlusion, viewpoint variation, and dense scenes. For the tracking subset, videos were converted into image sequences and annotated frame by frame in the ImageLabel tool, where both head and body regions were labeled as tracking targets. The resulting annotations were organized in the VOT [26] format, emphasizing temporal continuity and target consistency under occlusion, overlap, and high inter-individual similarity.

Segmentation tasks. For the instance segmentation subset, annotations were created using LabelMe[42], where each goat instance was manually delineated and assigned a class label. All annotations were stored in COCO format. The annotation process emphasized accurate instance separation and high-quality mask boundaries in realistic scenes with occlusion and overlap. For the semantic segmentation subset, pixel-level annotations were generated using the Baidu EasyData platform. Images were manually labeled by semantic regions, with emphasis on clear category definitions and consistent region assignment, to support fine-grained scene understanding in complex farm environments.

Identification and generation. For individual identification, no additional manual annotations were applied. Instead, images of each goat were organized into separate identity folders. The subset was constructed to preserve sample diversity across poses, viewpoints, and time while maintaining the high visual similarity between individuals. For image generation, manually filter out samples affected by blur, low contrast, or strong visual obstruction. Rather than using explicit spatial annotations, the images were organized by scene type, coat color, and posture to support image generation, domain translation, and data augmentation.

All annotations follow a unified naming convention and hierarchical directory structure. The annotation team consisted of researchers with relevant expertise, and quality control was performed through cross-validation among annotators, automated format checking, and manual review of randomly sampled labels. Together, these procedures enhance annotation consistency, reliability, and reproducibility. Fig 1 presents visual examples of annotations for several representative tasks in DGV-Suite. For a detailed description of the dataset, please visit our project website: https://y-f-wei.github.io/DGV-Suite/.

3.3 Data statistics and characteristics

The pose estimation dataset comprises 5,268 images, each annotated with 17 anatomical keypoints that represent major body regions, such as the torso and limbs. Each keypoint is further associated with a visibility indicator, allowing explicit evaluation of occlusion-aware pose estimation. Keypoint visibility statistics are shown in Fig 2. Real-world challenges, including self-occlusion, inter-animal overlap, and viewpoint variation in farm environments, lead to highly non-uniform visibility distributions across keypoints from different anatomical regions.

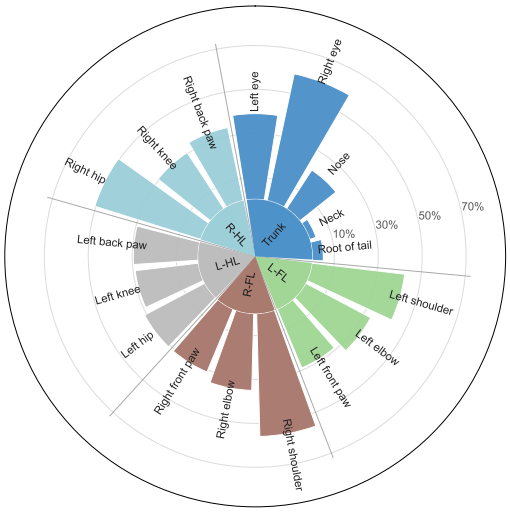

Figure 2: Percentage of invisible keypoints in the pose estimation subset.

The behavior recognition subset contains 1,259 video clips. It covers both indoor and outdoor farm environments and includes seven behavior categories, as shown in Fig 3a. Common daily behaviors such as standing, walking, and lying are represented with relatively sufficient samples. At the same time, the subset also includes rare but practically important abnormal behavior, such as paralyzing, which is difficult to collect in large quantities under real farming conditions; it preserves the natural frequency characteristics of real dairy goat behaviors. The combination of short and long clips further enables the benchmark to capture both transient and sustained motion patterns, making it suitable for benchmarking behavior recognition under realistic and diverse farm scenarios.

The object tracking subset includes 65 head-level sequences (45,915 frames) and 135 body-level sequences (115,285 frames), as shown in Fig 3b. All sequences were annotated frame by frame with temporally continuous labels and strict identity consistency, ensuring stable trajectories without erroneous ID switches. High-precision bounding boxes were maintained throughout the sequences. Collected from realistic farm scenes, the subset includes challenging cases such as occlusion, inter-animal overlap, strong appearance similarity, and dense group housing.

The object detection subset contains 3,430 images with 8,216 annotated instances. As shown in the outer ring of Fig 3c, it covers single-goat, multi-goat, and negative scenes, reflecting diverse real-world observations of farms. The substantial proportion of multi-target samples introduces realistic challenges, including occlusion, overlap, dense layouts, and pose variation, making the subset suitable for benchmarking robust detection models in complex dairy goat environments.

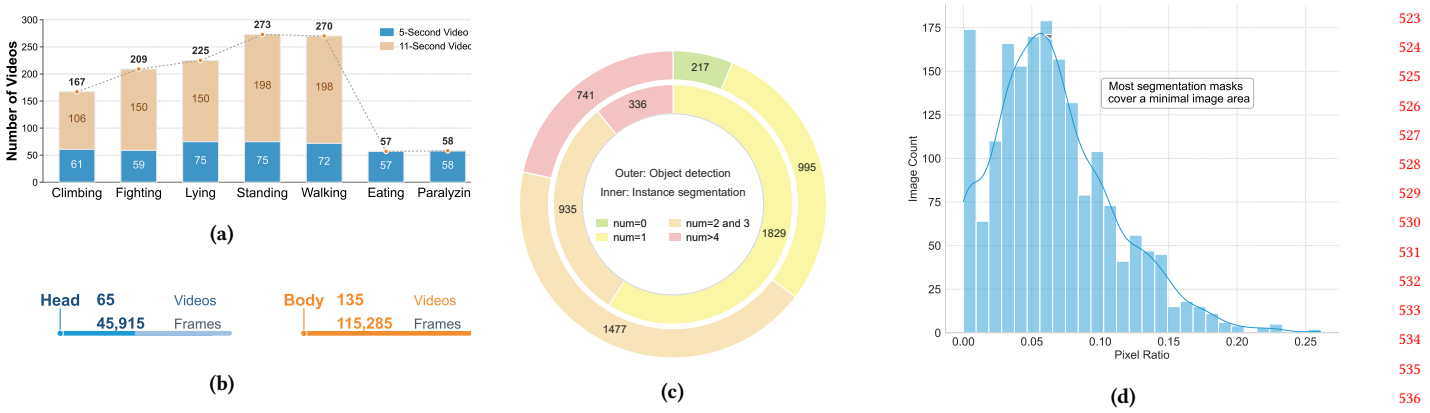

Figure 3: (a) Category distribution of the behavior recognition subset. (b) Annotations statistics of the object tracking subset. (c) Distribution of instance counts per image. The outer ring corresponds to object detection, and the inner ring corresponds to instance segmentation. (d) Distribution of mask annotations in the semantic segmentation subset.

The instance segmentation subset contains 3,100 images with 5,572 annotated instances. As shown in the inner ring of Fig 3c, it includes both isolated and crowded scenes with diverse poses and spatial interactions. A sufficient number of multi-instance samples provides challenging cases for precise mask prediction under occlusion, boundary ambiguity, and complex backgrounds, supporting fine-grained evaluation of instance segmentation models in realistic farm settings.

The semantic segmentation subset contains 1,829 images, each paired with a manually annotated pixel-level mask. The annotation protocol follows clear and unambiguous semantic category definitions, while emphasizing precise object boundaries and fine structural details. As shown in Fig 3d, the foreground pixel ratio varies across images, indicating substantial diversity in spatial coverage, scene composition, and target layout. Most masks occupy a relatively small portion of the image, which reflects realistic farm observations where goats often appear under partial views, complex poses, occlusion, and varied distances from the camera. Although most masks occupy a relatively small image area, the subset also includes denser foreground cases.

The individual identification subset contains 15,380 images from 159 dairy goats aged from 1 month to 8 years. This wide coverage of identities and age stages introduces substantial appearance variation in body size, shape, texture, and visual maturity, while preserving the high inter-individual similarity characteristic of dairy goats. In addition, the subset includes natural variations in pose, viewpoint, and capture conditions, making it suitable for evaluating identity-aware models under realistic cross-age and cross-scene settings.

The image generation subset contains 3,000 images organized by scene type (indoor/outdoor), color (black/white), and posture (standing/lying). This controlled yet diverse composition provides a practical basis for generative modeling, domain translation, and data augmentation. By covering multiple environmental and appearance factors, the subset supports the learning of generation models that preserve biologically plausible goat structure while adapting to diverse visual conditions.

4 Benchmark protocols

4.1 Tasks, splits, and evaluation metrics

We establish benchmark protocols on DGV-Suite across six supervised tasks, including object detection, instance segmentation, pose estimation, semantic segmentation, behavior recognition, and object tracking. These tasks are selected because they cover representative perception problems in dairy goat visual understanding and have well-defined supervised evaluation settings. For each task-specific subset, the data are divided into training, validation, and test sets with a ratio of 7:1:2. Model checkpoints are selected according to validation performance, early stopping is applied when the validation metric no longer improves, and all final results are reported on the test set.

Task-specific evaluation metrics follow standard practice in the corresponding research communities. For object detection and instance segmentation, we report AP, AP50, and AP75 to evaluate localization and mask quality under different overlap thresholds. For pose estimation, we also report AP, AP50, and AP75 to measure keypoint localization performance. For semantic segmentation, we adopt mIoU, mDice, and mPA to assess pixel-level region prediction quality. For behavior recognition, we report mAP, mAcc, and F1 to reflect classification performance under diverse behavioral categories. For object tracking, we use EAO, Success, and Speed to evaluate tracking accuracy, robustness, and efficiency.

4.2 Benchmark results

Table 2 summarizes the benchmark results of representative baselines on DGV-Suite across six supervised tasks. The underlined models are trained on 4060Ti and the rest on A40 GPU. Overall, more recent architectures generally outperform earlier designs, indicating the effectiveness of modern model families on dairy goat visual understanding. At the same time, the remaining performance gap across tasks highlights the practical difficulty and benchmark value of DGV-Suite especially under realistic farm conditions involving occlusion, dense animal layouts, strong inter-individual similarity, and complex backgrounds.

Table 2: Benchmark results of representative baselines on DGV-Suite across six supervised tasks

Object Detection				
Model	Backbone	AP	AP50	AP75
Faster R-CNN[38]	r50	0.5555	0.9845	0.6819
Cascade R-CNN[6]	r50	0.5930	0.9765	0.6689
FCOS[47]	r50	0.5160	0.9430	0.5090
YOLO v7[49]	E-ELAN	0.5994	**0.9878**	0.7936
YOLO v11[25]	C2PSA	**0.6346**	0.9826	**0.8023**
Instance Segmentation				
Mask R-CNN[21]	r50	0.7913	0.9775	0.9525
YOLACT[5]	r50	0.8107	0.9688	0.9288
SOLOv2[54]	r50	0.8707	0.9899	0.9792
YOLOv8-Seg[23]	CSPDarknet	0.6112	0.7991	0.6708
YOLOv12-Seg[46]	R-ELAN	**0.8930**	**0.9940**	**0.9884**
Pose Estimation				
HRNet[43]	w32	0.9139	**0.9766**	0.9454
PVT v2[52]	b2	0.9055	0.9750	0.9521
RTMPose[22]	CSPNeXt	**0.9145**	0.9764	0.9448
GRMPose[8]	CSPNeXt	0.9117	0.9765	**0.9550**
Semantic Segmentation				
Model	Backbone	mIoU	mDice	mPA
U-Net[41]	r50	0.7462	0.7983	0.7262
DeepLabV3+[9]	r50	**0.7750**	**0.8692**	**0.8618**
SegFormer[57]	MiT	0.7698	0.8638	0.8597
SegMAN[16]	SegMAN Encoder	0.7670	0.8616	0.8578
Behavior Recognition				
Model	Backbone	mAP	mAcc	F1
I3D[7]	r50	0.7690	0.7019	0.6890
SlowFast[14]	r50	0.7103	0.6935	0.6829
TimeSformer[3]	ViT	0.7710	0.7026	0.6890
VideoMAE v2[51]	ViT	0.8091	**0.7294**	**0.7256**
ELSlowFast-LSTM[59]	r50	**0.8195**	0.7195	0.7012
Object Tracking				
Model	Backbone	EAO	Success	Speed
SiamRPN++[28]	r50	0.2489	0.6401	33fps
STARK[58]	r50	0.2681	0.8037	25fps
SeqTrack[10]	ViT	0.2732	0.7945	21fps
ODTrack[60]	ViT	**0.2893**	**0.8102**	23fps

In object detection, YOLOv11[25] achieves the highest precision, surpassing both two-stage and anchor-free detectors. For instance segmentation, YOLOv12-Seg[46] attains the best performance, outperforming both two-stage frameworks and earlier one-stage models, while demonstrating strong robustness to illumination and occlusion. RTMPose[22], which is based on the CSPNeXt backbone, strikes the best balance between accuracy and efficiency in pose estimation. GRMPose[8], on the other hand, achieves higher precision when the evaluation thresholds are stricter. For semantic segmentation, DeepLabV3+[9] remains competitive through its efficient boundary refinement mechanism, outperforming transformer-based and multi-scale fusion models. In behavior recognition, EL SlowFast-LSTM[59] achieves superior temporal consistency, surpassing both CNN-based and transformer-based methods. Finally, in object tracking, ODTrack[60] reaches the highest EAO, benefiting from global feature interaction and faster convergence compared with prior transformer-based trackers.

5 Discussion

Benchmark difficulty and real-world relevance. The benchmark results demonstrate that DGV-Suite provides a practically challenging testbed for dairy goat visual understanding and that the dataset provides sufficient complexity to distinguish model capabilities while preserving practical relevance for dairy goat visual understanding. The difficulty mainly arises from realistic farm conditions, including dense layouts, inter-animal overlap, partial occlusion, strong appearance similarity, and complex indoor-outdoor backgrounds. For object detection and instance segmentation, performance is affected by dense layouts, inter-animal overlap, and frequent partial occlusion. For pose estimation, the non-uniform visibility distribution of anatomical keypoints reveals strong self-occlusion and viewpoint difficulty. For semantic segmentation, small object regions, irregular boundaries, and complex backgrounds make accurate region parsing challenging. For behavior recognition, the coexistence of short-term and long-term actions and category imbalance increases the difficulty of robust temporal modeling. For object tracking, long-term occlusion, close interactions, and strong appearance similarity between individuals make identity preservation particularly difficult.

Implications for model development. The benchmark results suggest that robust dairy goat visual understanding requires models that can jointly capture fine-grained local structure and broader contextual information. Tasks such as pose estimation and segmentation are sensitive to boundary quality, part-level details, and occlusion, while tracking and behavior recognition additionally require temporal continuity and identity-aware representation. In addition, no single architectural paradigm is consistently superior across all tasks, indicating that model effectiveness depends strongly on the match between task characteristics and architectural inductive biases. Therefore, DGV-Suite can serve not only as a benchmark for performance evaluation but also as a practical testbed for studying task-specific model design in precision livestock farming.

6 Conclusion

In this paper, we introduced DGV-Suite a large-scale benchmark suite of task-specific annotations for dairy goat vision. The dataset contains eight representative tasks in realistic intelligent farming scenarios. supports systematic evaluation for dairy goat visual understanding in authentic farm environments. We evaluated representative baselines on six supervised tasks. The results demonstrate the practical difficulty and benchmark value of DGV-Suite particularly under realistic challenges such as occlusion, dense layouts, high inter-individual similarity, and complex backgrounds. This also fills an important gap in dairy goat visual understanding, providing a practical resource for future research in precision livestock farming and intelligent multimedia analysis.

References

[1] Sanabel Abu Jwade, Andrew Guzzomi, and Ajmal Mian. 2019. On Farm Automatic Sheep Breed Classification Using Deep Learning. *Computers and Electronics in Agriculture* 167 (2019), 105055.

[2] William Andrew, Colin Greatwood, and Tilo Burghardt. 2017. Visual Localisation and Individual Identification of Holstein Friesian Cattle via Deep Learning. In *ICCVW*. 2850–2859.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *ICML*.

[4] Anil Bhujel, Yibin Wang, Yuzhen Lu, Daniel Morris, and Mukesh Dangol. 2025. A Systematic Survey of Public Computer Vision Datasets for Precision Livestock Farming. *Computers and Electronics in Agriculture* 229 (2025), 109718.

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. 2019. YOLACT: Real-Time Instance Segmentation. In *ICCV*. 9156–9165.

[6] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*. 6154–6162.

[7] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*. 4724–4733.

[8] Ling Chen, Lianyue Zhang, Jinglei Tang, Chao Tang, Rui An, Ruizi Han, and Yiyang Zhang. 2024. GRMPose: GCN-based Real-Time Dairy Goat Pose Estimation. *Computers and Electronics in Agriculture* 218 (2024), 108662.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*.

[10] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *CVPR*. 14572–14581.

[11] Felix Danso, Lukman Iddrisu, Shera Elizabeth Lungu, Guangxian Zhou, and Xianghong Ju. 2024. Effects of Heat Stress on Goat Production and Mitigating Strategies: A Review. *Animals* 14 (2024), 1793.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*. 248–255.

[13] C. Ducrot, M. B. Barrio, A. Boissy, F. Charrier, S. Even, P. Mormède, S. Petit, M. H. Pinard-van der laan, F. Schelcher, F. Casabianca, A. Ducos, G. Foucras, R. Guatteo, J. L. Peyraud, M. Vayssier-Taussat, P. Veysset, N. C. Friggens, and X. Fernandez. 2024. Animal Board Invited Review: Improving Animal Health and Welfare in the Transition of Livestock Farming Systems: Towards Social Acceptability and Sustainability. *animal* 18 (2024), 101100.

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *ICCV*. 6202–6211.

[15] Food and Agriculture Organization of the United Nations. 2024. Crops and livestock products. FAO Statistics Division. https://www.fao.org/faostat/en/#data/QCL

[16] Yunxiang Fu, Meng Lou, and Yizhou Yu. 2025. SegMAN: Omni-scale Context Modeling with State Space Models and Local Attention for Semantic Segmentation. In *CVPR*.

[17] Sigfredo Fuentes, Claudia Gonzalez Viejo, Eden Tongson, and Frank R. Dunshea. 2022. The Livestock Farming Digital Transformation: Implementation of New and Emerging Technologies Using Artificial Intelligence. *Animal Health Research Reviews* 23 (2022), 59–71.

[18] Yue Gao, Kai Yan, Baisheng Dai, Hongmin Sun, Yanling Yin, Runze Liu, and Weizheng Shen. 2023. Recognition of Aggressive Behavior of Group-Housed Pigs Based on CNN-GRU Hybrid Model with Spatio-Temporal Attention Mechanism. *Computers and Electronics in Agriculture* 205 (2023), 107606.

[19] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*.

[20] Liang Han, Pin Tao, and Ralph R. Martin. 2019. Livestock Detection in Aerial Images Using a Fully Convolutional Network. *Computational Visual Media* 5, 2 (2019), 221–228.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *ICCV*. 2980–2988.

[22] Tao Jiang, Peng Lu, Li Zhang, Ning Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399* (2023).

[23] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics

[24] Nathan A. Kelly, Bilal M. Khan, Muhammad Y. Ayub, Abir J. Hussain, Khalil Dajani, Yunfei Hou, and Wasiq Khan. 2024. Video Dataset of Sheep Activity for Animal Behavioral Analysis via Deep Learning. *Data in Brief* 52 (2024), 110027.

[25] Rahima Khanam and Muhammad Hussain. 2024. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint arXiv:2410.17725* (2024).

[26] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. 2016. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 11 (Nov 2016), 2137–2155.

[27] Timur Ledyaev, Konstantin Kolotyrin, Margarita Zabelina, Marianna Vasilchenko, and Diana Derunova. 2025. Main Aspects of Forming a Production Strategy for the Development of Dairy Goat Breeding. *Scientific Papers Series : Management, Economic Engineering in Agriculture and Rural Development* 25 (2025), 567–576.

[28] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*. 4282–4291.

[29] Xiangyuan Li, Cheng Cai, Ruifei Zhang, Lie Ju, and Jinrong He. 2019. Deep Cascaded Convolutional Models for Cattle Pose Estimation. *Computers and Electronics in Agriculture* 164 (2019), 104885.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*. Vol. 8693. N/A, 740–755.

[31] Cesar A. Meza-Herrera, Cayetano Navarrete-Molina, Ulises Macias-Cruz, Gerardo Arellano-Rodriguez, Angeles De Santiago-Miramontes, Maria A. Sariñana-Navarrete, Ruben I. Marin-Tinoco, and Carlos C. Perez-Marin. 2024. Dairy Goat Production Systems: A Comprehensive Analysis to Reframe Their Global Diversity. *Animals* 14 (2024), 3717.

[32] Beth A. Miller and Christopher D. Lu. 2019. Current Status of Global Dairy Goat Production: An Overview. *Asian-Australasian Journal of Animal Sciences* 32 (2019), 1219–1232.

[33] Suresh Neethirajan. 2020. The Role of Sensors, Big Data and Machine Learning in Modern Animal Farming. *Sensing and Bio-Sensing Research* 29 (2020), 100367.

[34] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. 2022. Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding. In *CVPR*.

[35] Kian Eng Ong, Sivaji Retta, Ramarajulu Srinivasan, Shawn Tan, and Jun Liu. 2023. CattleEyeView: A Multi-Task Top-down View Cattle Dataset for Smarter Precision Livestock Farming.

[36] Ahmed Qazi, Taha Razzaq, and Asim Iqbal. 2024. AnimalFormer: Multimodal Vision Framework for Behavior-Based Precision Livestock Farming. In *CVPR*. 7973–7982.

[37] Yongliang Qiao, Cameron Clark, Sabrina Lomax, He Kong, Daobilige Su, and Salah Sukkarieh. 2021. Automated Individual Cattle Identification Using Video Data: A Unified Deep Learning Architecture Approach. *Frontiers in Animal Science* 2 (2021).

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[39] Martin Riekert, Achim Klein, Felix Adrion, Christa Hoffmann, and Eva Gallmann. 2020. Automatically Detecting Pig Position and Posture by 2D Camera Imaging and Deep Learning. *Computers and Electronics in Agriculture* 174 (2020), 105391.

[40] Ali Rohan, Muhammad Saad Rafaq, Md. Junayed Hasan, Furqan Asghar, Ali Kashif Bashir, and Tania Dottorini. 2024. Application of Deep Learning for Livestock Behaviour Recognition: A Systematic Literature Review. *Computers and Electronics in Agriculture* 224 (2024), 109115.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-assisted Intervention (MICCAI)*. 234–241.

[42] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision* 77 (2008), 157–173.

[43] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*. 5686–5696.

[44] Qianqian Sun, Shuqin Yang, Meili Wang, Shenrong Hu, and Jifeng Ning. 2024. A Real-Time Dairy Goat Tracking Based on MixFormer with Adaptive Token Elimination and Efficient Appearance Update. *Computers and Electronics in Agriculture* 218 (2024), 108645.

[45] Jinglei Tang, Guoxin Yang, Yurou Sun, Jing Xin, and Dongjian He. 2019. Salient Object Detection of Dairy Goats in Farm Image Based on Background and Foreground Priors. *Neurocomputing* 332 (2019), 270–282.

[46] Yunjie Tian, Qixiang Ye, and David Doermann. 2025. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv preprint arXiv:2502.12524* (2025).

[47] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *ICCV*. 9626–9635.

[48] Jehan-Antoine Vayssade, Rémy Arquet, and Mathieu Bonneau. 2019. Automatic Activity Tracking of Goats Using Drone Camera. *Computers and Electronics in Agriculture* 162 (2019), 767–772.

[49] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *CVPR*. 7464–7475.

[50] Dong Wang, JingLei Tang, Weijie Zhu, Huan Li, Jing Xin, and Dongjian He. 2018. Dairy Goat Detection Based on Faster R-CNN from Surveillance Video. *Computers and Electronics in Agriculture* 154 (2018), 443–449.

[51] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *CVPR*. 14549–14560.

[52] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2022. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* 8, 3 (2022), 415–424.

[53] Xiaobo Wang, Yufan Hu, Meili Wang, Mei Li, Wenxiao Zhao, and Rui Mao. 2024. A Real-Time Lightweight Behavior Recognition Model for Multiple Dairy Goats. *Animals* 14 (2024), 3667.

[54] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. 2020. SOLOv2: dynamic and fast instance segmentation. In *NeurIPS*. Article 1487, 12 pages.

[55] Jason V. Watters, Bethany L. Krebs, and Caitlin L. Eschmann. 2021. Assessing Animal Welfare with Behavior: Onward with Caution. *Journal of Zoological and Botanical Gardens* 2 (2021), 75–87.

[56] Zachary Winkler, Laura E. Boucheron, Santiago Utsumi, Shelemia Nyamuryekung'e, Matthew McIntosh, and Richard E. Estell. 2024. Effects of Dataset Curation on Body Condition Score (BCS) Determination with a Vision Transformer (ViT) Applied to RGB+depth Images. *Smart Agricultural Technology* 8 (2024), 100482.

[57] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*.

[58] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. 2021. Learning spatio-temporal transformer for visual tracking. In *ICCV*. 10448–10457.

[59] Junpeng Zhang, Zihan Bai, Yifan Wei, Jinglei Tang, Ruizi Han, and Jiaying Jiang. 2025. Behavior Detection of Dairy Goat Based on YOLO11 and ELSlowFast-LSTM. *Computers and Electronics in Agriculture* 234 (2025), 110224.

[60] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. 2024. ODTrack: Online Dense Temporal Token Learning for Visual Tracking. In *AAAI*, Vol. 38. 7588–7596.

[61] Ali Zia, Renuka Sharma, Reza Arablouei, Greg Bishop-Hurley, Jody McNally, Neil Bagnall, Vivien Rolland, Brano Kusy, Lars Petersson, and Aaron Ingham. 2023. CVB: A Video Dataset of Cattle Visual Behaviors. arXiv:2305.16555 [cs]