

Improving the Reliability of Ear-Tip Recognition in a Trap: From Ear Counting to Structural Decision Conditions

朝陽科技大學 [學號] [姓名]

[組員 2 / 學校 / 學號 / 姓名]

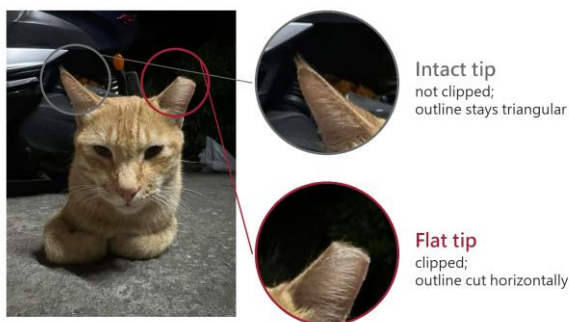
[組員 3]

[組員 4]

Abstract

Trap-Neuter-Return (TNR) marks neutered cats by clipping one ear tip. We have built a trap prototype that runs ear-tip recognition on a microcontroller and decides on its own whether to close the gate. Its current logic accumulates clipped and unclipped detection boxes over an observation window and captures according to their ratio. Field deployment exposed three limitations: it counts detection boxes rather than frames, so one pose is counted repeatedly, and applies no structural check; the model was trained in a single stage without fine-tuning on field data; and clipped-ear samples are structurally scarce, capping attainable accuracy. This project addresses all three, with two-stage transfer learning, training-time augmentation, and five structural conditions replacing the count comparison, so that the system can separate “the image is insufficient for a decision” from “the image is sufficient but no clipped ear was seen”. We expect a higher share of valid evidence frames and a traceable basis for every decision. A negative result would also be informative: if evidence remains insufficient, the bottleneck is the camera placement, not the model.

1. Introduction



Same cat: left ear intact, right ear clipped (TNR clips only one tip)

Figure 1. Ear-tip clipping, photographed in the field. TNR clips only one tip, so a neutered cat always carries one intact ear and one flat-tipped ear. The difference is a few centimetres of outline, and recognising it automatically is the task of this project.

Motivation. Since Taiwan stopped euthanasia for population control in 2017, shelter capacity has been the binding constraint [1], so reducing reproduction became the main route. TNR has the most complete literature and the clearest measured effect on population size [2]; it works only if un-neutered individuals are caught, and the sole

field cue is the clipped ear tip (Figure 1) [3]. A conventional trap merely triggers a gate, so every animal that enters is caught. Where TNR runs continuously the neutered share accumulates; at our study site about 80% of individuals are already neutered, so most captures are recaptures, costing labour and causing needless stress. The chain-driven gate is also loud, so nearby cats may learn to avoid the trap.

Problem definition. A single YOLO model emits cat-face, clipped-ear and unclipped-ear boxes in each frame. The state machine enters observation on the first frame containing a face, accumulates the two ear counts over the window, and at the end judges the animal un-neutered and closes the gate if the unclipped count exceeds a fixed multiple of the clipped count. This works on clear images but shows three limitations in the field. First, it accumulates boxes, not frames: a pose held over consecutive frames is counted repeatedly, so one piece of evidence votes many times. Second, it applies no structural check: it never verifies that exactly one face and two ears are present, that the ear boxes lie inside the face box, or that both ears are not clipped, which cannot occur under TNR. A side view or one ear boxed twice can still satisfy the count, so a decision is made on insufficient evidence and nothing records what it saw. Third, because TNR clips only one tip, clipped and unclipped ears always co-occur, so clipped samples, the sole evidence of neutering, are structurally scarce.

Existing approaches and their gap. Smart-trap research targets species-level discrimination: the AI box trap of Besala et al. decides whether the animal belongs to the target species [5]. Our two classes are two states of the same species, separated by a few centimetres of ear outline (Figure 1). Edge-versus-cloud placement has been studied for latency and energy [4], but that work assumes the detector output can be acted upon directly and does not ask how a device should behave when its evidence is insufficient. Raising average precision alone is not enough: the system must also judge whether a frame is qualified to support a decision.

Scope. The hardware, firmware and gate mechanism are already built and running in the field. This project changes only how the model is trained and how a detection leads to a capture decision, reusing the existing device and data; the limitations above were confirmed against the current code and training configuration.

Proposed approach. Section 2 details three measures answering the three limitations: two-stage transfer learning, training-time augmentation, and five structural conditions that mark a failing frame unusable instead of feeding it into the decision.

2. Method

2.1. Two-stage transfer learning. Stage one starts from COCO weights and learns general cat-face and ear features from about 1,700 general cat images. Stage two starts from those weights and fine-tunes on about 600 field images taken inside the trap, adapting the model to the deployed angle, lighting and distance. Field data alone are too few for a stable representation, and general data contain no in-cage conditions, so the two must be used in sequence. The model is YOLOv7-tiny at 640×640 input, matching the on-device accelerator.

2.2. Data augmentation. Augmentation runs on the fly during training, so duplicated samples do not consume dataset volume. Planned settings are Mosaic four-image tiling (1.0), horizontal flip (0.5), HSV jitter ($H 0.015 / S 0.7 / V 0.4$), translation and scaling (0.1 / 0.5), and small MixUp and Paste-in (0.05 each). Vertical flip and rotation are excluded, as they do not occur in the cage. Per-class precision before and after will show whether the clipped-ear class benefits.

2.3. Structural decision conditions. Five conditions replace the count comparison; a frame must pass all of them to count as valid evidence: (i) exactly one cat face; (ii) exactly two ears; (iii) not both ears clipped; (iv) both ear boxes overlapping the face box; (v) each of the three class confidences above threshold. Condition (iii) follows from TNR clipping only one tip: two clipped ears never occur, so they indicate a classification error. Conditions (i), (ii) and (iv) test image structure and are independent of thresholds, which lets the system distinguish two kinds of failure and makes each reason countable. The accumulation unit also changes from detection boxes to valid frames, with a minimum counting interval.

2.4. Evidence accumulation. Valid frames enter one of two parallel tracks by clipped-ear confidence: the high track (≥ 75) needs 2 frames to declare the animal neutered, the low track (≥ 65) needs 3. Running both lets the system finish early on clear images while retaining capability on ordinary ones. Valid frames must be at least 1 s apart. All thresholds sit in one configuration file for later tuning.

2.5. Evaluation. Every observation window will be logged frame by frame: the inference result, which structural conditions passed, and the thresholds in force. Three indicators follow: the share of valid evidence frames, the distribution of rejections across conditions, and per-class recall. The second is decisive: rejections dominated by the confidence threshold point to model precision, those dominated by structural conditions point to camera placement, and the two demand entirely different remedies.

2.6. Deployment constraint. Every improvement must run on the existing RTL8735B microcontroller: the model must quantise into the on-device format, and the decision logic must fit the existing firmware structure. Any approach failing this is excluded, however well it performs offline.

3. Expected Results

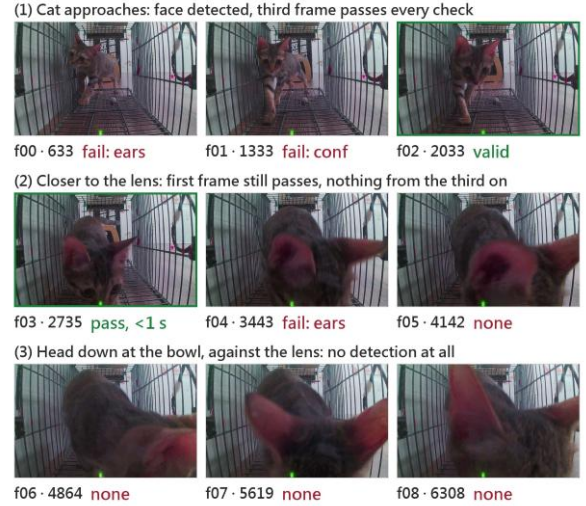


Figure 2. One observation window recorded in the field (nine frames; times in milliseconds from the start of observation). Only one frame becomes valid evidence; the rest either fail a structural condition or yield no detection at all once the cat is against the lens. The current rule cannot separate these two cases; the structural check is expected to mark the former and record which condition failed.

Aspect	Current rule	Target of this project
Decision basis	ratio of ear box counts	five structural conditions, two tracks
Accumulation unit	detection boxes	valid frames, throttled
Training	single stage	two-stage transfer learning
Unusable images	not distinguished	marked, with reason recorded

Table 1. The current rule against the targets of this project.

We report per-class recall, not overall accuracy: where about 80% of individuals are already neutered, an “always neutered” rule alone scores about 80%, which says nothing. Three outputs are expected: the gain in the share of valid evidence frames, the distribution of rejections across structural conditions, and the change in clipped-ear precision before and after two-stage training and augmentation. A negative result remains possible; even then the rejection distribution locates the bottleneck, and separating and quantifying these two causes is the substantive contribution of this project.

References

- [1] Yan, T.-Y., and Teng, K. T.-y. Trends in animal shelter management, adoption, and animal death in Taiwan from 2012 to 2020. *Animals*, 13:1451, 2023.

- [2] Smith, L. M., Hartmann, S., Munteanu, A. M., Dalla Villa, P., Quinnell, R. J., and Collins, L. M. The effectiveness of dog population management: A systematic review. *Animals*, 9:1020, 2019.
- [3] Agriculture Knowledge Portal, Taiwan. Introduction to TNR. 2020.
- [4] Koubaa, A., Ammar, A., Alahdab, M., Kanhouh, A., and Azar, A. T. DeepBrain: Experimental evaluation of cloud-based computation offloading and edge computing in the Internet-of-Drones for deep learning applications. *Sensors*, 20:5240, 2020.
- [5] Besala, F. I., Niimoto, R., Lee, J. H., and Okamoto, S. Development of AI-based smart box trap system for capturing a harmful wild boar. *ROBOMECH Journal*, 12:4, 2025.