

الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08 ماي 1945، بسبدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès



THESIS

To obtain the diploma of **Engineer**
Field: **Computer Science**
Specialty: **Intelligence Artificielle et Science des Données**

Theme

Adversarial Embedding at Scale: A Verified Implementation of the Backpack Min–Max Protocol and Its Behaviour Across the Spatial and JPEG Domains

Presented by:
Aymen Slimani

Submission Date: **September 2026**
In front of the jury composed of:

Dr. Taouli Amina
Dr. Amrane Abdelkader
Dr. Louhab Zouaoui

President
Supervisor
Examiner

The Backpack scheme of Bernard *et al.* (IEEE TIFS, 2022) makes adversarial steganographic embedding fully backpropagable: it replaces the two non-differentiable links in the min–max protocol (the discrete embedding change and the payload constraint) by a Gumbel-softmax relaxation and an implicitly differentiated Lagrange multiplier. The companion master’s report develops that theory. **This engineering report is the practical half: it reports what happens when the scheme is rebuilt from scratch, verified equation by equation, and run at full scale on a production GPU cluster.**

The work is organised as a staged ablation along three axes (binary before ternary embedding, Fashion-MNIST before BOSSBase, spatial before JPEG) so that every regression is attributable to a single component. Three results follow. *First*, in the JPEG domain the reproduction succeeds: on BOSSBase at QF75 and 0.4bpnzAC against a retrained XuNet, the detector’s probability of error climbs from $P_{\text{err}} = 0.0947$ at the J-UNIWARD baseline to a peak of $P_{\text{err}} = 0.4494$ after seven protocol iterations, against the $0.075 \rightarrow 0.476$ reported in the source paper. *Second*, the protocol was made roughly $36\times$ faster per cover through a batched attack, a safeguarded-Newton λ solver, argmax detector routing and dynamic batch sizing, every one of them gated behind a lossless-equivalence test suite that proves the mathematics is unchanged. *Third*, in the spatial domain on Fashion-MNIST the protocol improves the embedding at every iteration, raising P_{err} by roughly five points and more than tripling the attack success rate, but at a rate an order of magnitude below the JPEG domain. Per-cover telemetry over 10,000 covers localises the cause: between 66% and 90% of covers exhaust the optimisation budget without reaching the detector’s decision threshold, and the median such cover stops a full logit short of it. Two candidate explanations were eliminated by measurement along the way, a mis-specified Monte-Carlo sample schedule and detectors that are provably blind at low payload, and two further hypotheses were refuted directly: temperature annealing is confirmed working, and threshold overshoot, which is real in the JPEG domain, is shown not to apply in the spatial one. An incorrect definition of attack success, measured against the cover’s own score rather than the detector’s decision threshold, was identified and corrected.

Finally, a controlled four-arm race under an automated fairness audit compares the paper’s Softmax-Gumbel estimator against ReinMax, Gumbel-Rao and RLOO inside the protocol.

Keywords: adversarial embedding, steganography, min–max protocol, Backpack, Gumbel-softmax, gradient estimators, reproduction study, GPU cluster engineering, XuNet, BOSSBase.