Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών
Προηγμένα Θέματα Τεχνητής Νοημοσύνης
9ο εξάμηνο

Benchmarking Monocular Depth Estimation Models

Υπεύθυνοι εργασίας:
Α. Βουλόδημος
Β. Καραμπίνης
Η. Μήτσουρας

Νικόλας Σπυρόπουλος, ΑΜ: 03121202

1 Περίληψη

Η κατανόηση της τρισδιάστατης δομής ενός σκηνικού αποτελεί θεμελιώδη στόχο του τομέα της Όρασης Υπολογιστή (Computer Vision), με εφαρμογές στην αυτόνομη οδήγηση, τη ρομποτική πλοήγηση, την επαυξημένη πραγματικότητα (AR), την τρισδιάστατη ανακατασκευή και την ανάλυση σκηνών. Ένα από τα πιο σημαντικά προβλήματα που συνδέονται με αυτή την κατανόηση είναι η **εκτίμηση βάθους (Depth Estimation)**, δηλαδή ο υπολογισμός της απόστασης κάθε σημείου της εικόνας (pixel) από την κάμερα.

Τα τελευταία χρόνια, έχουν αναπτυχθεί μοντέλα εκτίμησης βάθους που μπορούν να παράγουν ακριβείς χάρτες βάθους ακόμη και από μία μόνο RGB εικόνα. Η συγκεκριμένη εφαρμογή ονομάζεται **Monocular Depth Estimation** καθώς δεν χρειάζεται δεδομένα από πολλούς αισθητήρες ή πολλές κάμερες, παρά μόνο μια εικόνα. Στο πλαίσιο αυτής της εργασίας, εξετάζονται τέσσερα σύγχρονα μοντέλα —**MiDaS**, **ZoeDepth**, **Depth-Anything-V2** και **Marigold**— τα οποία αντιπροσωπεύουν διαφορετικές αρχιτεκτονικές, τεχνικές μάθησης και επίπεδα ικανότητας στη γενίκευση μεταξύ τομέων (**domain generalization**). Αφού πρώτα, αναλυθούν οι αρχιτεκτονικές και οι ιδιαιτερότητες τους, θα γίνει σύγκριση της ακρίβειας των μοντέλων χρησιμοποιώντας ευρέως διαδεδομένα σύνολα δεδομένων(datasets).

Πριν παρουσιαστούν τα μοντέλα, είναι απαραίτητο να δοθούν οι βασικές έννοιες της εκτίμησης βάθους και τα είδη της, καθώς και το πρόβλημα της γενίκευσης τομέα, το οποίο αποτελεί κρίσιμο σημείο για την επιτυχία των μοντέλων Monocular Depth Estimation.

2 Εκτίμηση Βάθους (Depth Estimation)

Η εκτίμηση βάθους στοχεύει στην εξαγωγή ενός **χάρτη βάθους (depth map)**, στον οποίο για κάθε pixel αντιστοιχεί μια εκτίμηση της απόστασής του από την κάμερα. Παραδοσιακές προσεγγίσεις στηρίζονται σε στερεοσκοπικά συστήματα ή σε αισθητήρες ενεργού βάθους (όπως LiDAR ή ToF). Ωστόσο, η δυνατότητα πρόβλεψης βάθους από μόνο μια οπτική μιας εικόνας (monocular depth estimation) είναι ιδιαίτερα ελκυστική λόγω του χαμηλού κόστους, της ευκολίας ενσωμάτωσης και της ευρείας διαθεσιμότητας συστημάτων που διαθέτουν μόνο μία κάμερα.

Η εκτίμηση βάθους από μία μόνο εικόνα αποτελεί μη-σαφώς ορισμένο πρόβλημα (**ill-posed**), καθώς άπειρες τρισδιάστατες σκηνές μπορούν να προβάλουν την ίδια δισδιάστατη εικόνα. Γι' αυτό, τα σύγχρονα μοντέλα βασίζονται σε μεγάλο βαθμό στη μάθηση στατιστικών και γεωμετρικών προτύπων από δεδομένα [1].

Η εκτίμηση βάθους διακρίνεται σε δύο βασικές υποκατηγορίες[1]: **Απόλυτη (Absolute/Metric)** και **Σχετική (Relative)** εκτίμηση βάθους.

2.1 Απόλυτη Εκτίμηση Βάθους (Absolute Depth Estimation)

Η απόλυτη εκτίμηση βάθους αναφέρεται στην πρόβλεψη της πραγματικής απόστασης κάθε σημείου της σκηνής από την κάμερα, σε φυσικές μονάδες όπως μέτρα. Σε αυτή την περίπτωση, το μοντέλο πρέπει να εκτιμήσει όχι μόνο το σχήμα και τη δομή της σκηνής, αλλά και την κλίμακα της, ώστε ο χάρτης βάθους να ανταποκρίνεται ακριβώς στην πραγματικότητα. Η εκπαίδευση τέτοιων μοντέλων απαιτεί δεδομένα με μετρικές ετικέτες, συνήθως από αισθητήρες όπως LiDAR, structured light ή Time-of-Flight. Εξαιτίας αυτής της εξάρτησης από συγκεκριμένα συστήματα λήψης για ακριβείς μετρήσεις, η απόλυτη εκτίμηση βάθους είναι ευαίσθητη σε αλλαγές τομέα, όπως διαφορετικές κάμερες, φωτιστικές συνθήκες ή τύπους σκηνών. Παρ' όλα αυτά, αποτελεί αναντικατάστατη προσέγγιση σε εφαρμογές όπου η πραγματική απόσταση είναι κρίσιμη, όπως η αυτόνομη οδήγηση, η ρομποτική πλοήγηση και η τρισδιάστατη χαρτογράφηση.

2.2 Σχετική Εκτίμηση Βάθους (Relative Depth Estimation)

Η σχετική εκτίμηση βάθους επικεντρώνεται στην πρόβλεψη της δομής και της τοπολογίας της σκηνής, χωρίς να επιδιώκει να εκφράσει τις αποστάσεις σε πραγματικές μονάδες. Ο χάρτης βάθους που παράγει ένα τέτοιο μοντέλο είναι σε κλίμακα η οποία δεν έχει κάποια φυσική σημασία, αποτυπώνοντας όμως με συνέπεια το ποια σημεία βρίσκονται πιο κοντά ή πιο μακριά από την κάμερα. Επειδή δεν απαιτείται μετρικό ground truth, η σχετική εκτίμηση βάθους μπορεί να εκπαιδευτεί σε πολύ μεγάλα, ετερογενή σύνολα δεδομένων, συχνά προερχόμενα από διαφορετικές πηγές ή με ποικιλία σκηνών. Αυτό καθιστά τη μέθοδο ιδιαίτερα ανθεκτική σε αλλαγές τομέα, με αποτέλεσμα να επιτυγχάνει καλύτερη γενίκευση σε νέες συνθήκες, διαφορετικούς τύπους σκηνών ή διαφορετικούς αισθητήρες. Για τον λόγο αυτό, πολλά σύγχρονα μοντέλα που δίνουν έμφαση στην ευρεία γενίκευση υιοθετούν τη σχετική εκτίμηση βάθους, ιδιαίτερα όταν η ακριβής κλίμακα δεν αποτελεί προϋπόθεση για την εκάστοτε εφαρμογή.

[1] `https://huggingface.co/docs/transformers/tasks/monocular_depth_estimation`

3 Γενίκευση Τομέα (Domain Generalization)

Η γενίκευση τομέα αποτελεί ένα από τα βασικά ζητήματα στη εκτίμηση βάθους από μία εικόνα και αναφέρεται στην ικανότητα ενός μοντέλου να διατηρεί υψηλή απόδοση όταν εφαρμόζεται σε δεδομένα που διαφέρουν από αυτά στα οποία εκπαιδεύτηκε. Ένα μοντέλο που δεν γενικεύει καλά μπορεί να λειτουργεί άριστα σε ένα συγκεκριμένο σύνολο δεδομένων, αλλά να παρουσιάζει σημαντική πτώση στην απόδοση όταν αντιμετωπίζει διαφορετικές σκηνές, συνθήκες φωτισμού ή κάμερες.

Η αντιμετώπιση του προβλήματος της γενίκευσης τομέα απαιτεί ειδικές στρατηγικές, όπως η εκπαίδευση σε ποικιλόμορφα και πολυδιάστατα datasets, η χρήση τεχνικών κανονικοποίησης και scale-invariant loss functions [2], καθώς και η αξιοποίηση μεγάλων συνόλων συνθετικών [3] και πραγματικών δεδομένων. Όλα τα μοντέλα που εξετάζονται στην εργασία —MiDaS, ZoeDepth, Depth-Anything-V2 και Marigold— αντιμετωπίζουν το ζήτημα της γενίκευσης με διαφορετικούς τρόπους, στοχεύοντας στην επίτευξη αξιόπιστης απόδοσης σε ποικίλες συνθήκες και σκηνές, ανεξαρτήτως των δεδομένων εκπαίδευσης.

Αναφορές

[1] J. Zhang, *Survey on Monocular Metric Depth Estimation*, 2025. arXiv: 2501.11841 [cs.CV]. διεύθν.: https://arxiv.org/abs/2501.11841.

[2] D. Eigen, C. Puhrsch και R. Fergus, *Depth Map Prediction from a Single Image using a Multi-Scale Deep Network*, 2014. arXiv: 1406.2283 [cs.CV]. διεύθν.: https://arxiv.org/abs/1406.2283.

[3] Y. Yao κ.ά., "Improving Domain Generalization in Self-supervised Monocular Depth Estimation via Stabilized Adversarial Training", στο *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Νοέ. 2024, σσ. 183–201, ISBN: 9783031726910. DOI: 10.1007/978-3-031-72691-0_11. διεύθν.: http://dx.doi.org/10.1007/978-3-031-72691-0_11.