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.

4 MiDaS (Mixing Datasets for Zero-shot Cross-dataset Transfer)

Το MiDaS αντιπροσωπεύει μια κομβική εξέλιξη στην εκτίμηση βάθους από μία μόνο εικόνα, καθώς αντιμετωπίζει συστηματικά το κρίσιμο πρόβλημα της γενίκευσης τομέα. Η επιτυχία του μοντέλου έγκειται στην ανάπτυξη μεθοδολογιών για την αποτελεσματική **σύνθεση ετερογενών συνόλων δεδομένων (mixing datasets)**, ακόμη και αν οι αρχικές τους ετικέτες βάθους είναι ασύμβατες. Οι πρώτες δύο εκδόσεις του MiDaS χρησιμοποιούσαν συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNNS) [4], ενώ η τρίτη και τελευταία έκδοση υιοθετεί διαφορετικές αρχιτεκτονικές Vision Transfer (ViT) [5] ακολουθώντας τις εξελίξεις στο πεδίο της όρασης υπολογιστή.

4.1 Αρχιτεκτονική

Η αρχιτεκτονική του MiDaS εξελίχθηκε σε δύο κύριες φάσεις, διατηρώντας πάντα τη βασική δομή **Κωδικοποιητή-Αποκωδικοποιητή (Encoder-Decoder)**.

4.1.1 MiDaS v1 & v2 (CNN-based)

Οι αρχικές εκδόσεις του MiDaS βασίστηκαν στα **Συνελικτικά Νευρωνικά Δίκτυα (CNNs)**, τα οποία ήταν η κυρίαρχη τεχνολογία για την Όραση Υπολογιστή την εποχή της πρώτης δημοσίευσης [4].

- **Κωδικοποιητής (Encoder)**: Χρησιμοποιήθηκε ένα δίκτυο υψηλής χωρητικότητας τύπου **ResNeXt-101-WSL (Weakly-Supervised Learning)**. Η επιλογή ενός τόσο βαθιού δικτύου και η προ-εκπαίδευση σε μαζικά, αδύναμα επιβλεπόμενα δεδομένα (WSL) επέτρεψε στον κωδικοποιητή να μάθει εξαιρετικά ισχυρές και γενικεύσιμες αναπαραστάσεις χαρακτηριστικών (Transfer Learning), απαραίτητες για την επιτυχία του zero-shot transfer. Στην αρχική δημοσίευση αναφέρεται επίσης ότι και με χρήση ενός μικρότερου κωδικοποιητή, όπως το **ResNet-50**, το μοντέλο κατάφερε να επιτύχει καλύτερη απόδοση από τα τότε state-of-the-art μοντέλα.

- **Αποκωδικοποιητής / Κεφαλή προβλέψεων (Decoder / Prediction Head)**: Ο αποκωδικοποιητής ήταν υπεύθυνος για την ανασύνθεση του χάρτη βάθους από τα αφηρημένα χαρακτηριστικά του κωδικοποιητή. Χρησιμοποιούσε τεχνικές **multi-scale feature fusion**[2] μέσω συνδέσεων παράκαμψης (Skip Connections), για να συνδυάσει λεπτομερή χωρικά χαρακτηριστικά (από ρηχά στρώματα) με εννοιολογικές πληροφορίες (από βαθιά στρώματα), επαναφέροντας την ανάλυση στην αρχική διάσταση της εικόνας μέσω **ανοδικής δειγματοληψίας (Upsampling)**.

4.1.2 MiDaS v3.0 & v3.1 / DPT (Transformer-Based)

Οι νεότερες εκδόσεις του MiDaS (συχνά αναφερόμενες ως **DPT - Dense Prediction Transformer**) διατήρησαν τη μεθοδολογία εκπαίδευσης, αλλά αντικατέστησαν τον CNN κωδικοποιητή με ένα **Vision Transformer (ViT)** [5].

- **Κωδικοποιητής (Encoder)**: Υιοθετήθηκαν διάφοροι ViT backbones, όπως ViT, BEIT, Swin και SwinV2. Η επιτυχία τους οφείλεται στην ικανότητά τους να μοντελοποιούν μακροχρόνιες εξαρτήσεις (long-range dependencies) σε όλη την εικόνα, καταγράφοντας αποτελεσματικότερα την ολική γεωμετρία (global structure) της σκηνής. Αυτό βελτιώνει περαιτέρω τη γενίκευση σε σκηνές με άγνωστη δομή.

- **Αποκωδικοποιητής (Decoder)**: Η αρχιτεκτονική του αποκωδικοποιητή προσαρμόστηκε για να λαμβάνει ως είσοδο τα "**tokens**" (τα διακριτά

[2] `https://medium.com/@nbeel.original/getting-started-with-depth-estimation-using-midas-and-python-d0119bfe1159`

τμήματα πληροφορίας) από τα διάφορα στάδια του Transformer. Στην συνέχεια "συναρμολογεί" εκ νέου (**reassemble**) αυτά τα tokens σε αναπαραστάσεις εικόνας πολλαπλών αναλύσεων, τις οποίες επεξεργάζεται για να παραχθεί ο τελικός πυκνός χάρτης βάθους.

4.2 Εκπαίδευση (Training)

Η στρατηγική εκπαίδευσης του MiDaS είναι η κύρια καινοτομία του και παραμένει σταθερή σε όλες τις εκδόσεις, εξασφαλίζοντας την αρχιτεκτονική ανεξαρτησία της μεθόδου. Σχεδιάστηκε για να επιτύχει **Zero-shot Cross-dataset Transfer**, δηλαδή το μοντέλο αξιολογείται σε ένα test dataset το οποίο δεν είναι υποσύνολο του train dataset.

Το μοντέλο εκπαιδεύτηκε συνδυάζοντας δεδομένα από **πολλαπλά και ετερογενή σύνολα (datasets)**, όπως ReDWeb, MegaDepth, DIML Indoor, WSVD και, κυρίως, μια νέα, μαζική πηγή από καρέ 3D ταινιών (3D Movies). Αυτή η ανάμειξη εξασφαλίζει ότι το μοντέλο εκτίθεται σε τεράστια ποικιλία σκηνών (πχ Indoor/Outdoor, Static/Dynamic), καθιστώντας τις αναπαραστάσεις που προσπαθεί να μάθει ανθεκτικές και γενικές. Χρησιμοποιήθηκε η τεχνική **Pareto-optimal Multi-objective Optimization (Βελτιστοποίηση Πολλαπλών Στόχων)** η οποία αντιμετωπίζει την εκπαίδευση σε κάθε dataset ως ξεχωριστό στόχο, εξασφαλίζοντας ισορροπημένη μάθηση ώστε η βελτίωση της απόδοσης σε ένα dataset να μην υποβαθμίζει την απόδοση σε κάποιο άλλο (φαινόμενο που παρατηρήθηκε σε πείραμα εκπαίδευσης με μια **αφελή - naive μέθοδο**).

Για να καταστεί δυνατή η συνεκπαίδευση σε datasets με ασύμβατες ετικέτες (πχ μετρικό βάθος έναντι σχετικού βάθους), το MiDaS παράγει τις προβλέψεις του στον χώρο της **Δυσαναλογίας (Disparity Space)** (δηλαδή του αντίστροφου βάθους, D^{-1}), ο οποίος είναι αριθμητικά σταθερός και συμβατός με τις όλες τις πηγές των ground truths σε όλα τα σύνολα δεδομένων. Για να το καταφέρει αυτό εισάγει μια καινοτόμο συνάρτηση απώλειας (Loss function). Χρησιμοποιείται η **Scale- and Shift-Invariant Trimmed MAE** ($\mathcal{L}_{ssitrim}$) η οποία είναι αδιάφορη ως προς την κλίμακα (s) και τη μετατόπιση (t) (scale- and shift-invariant), επιτρέποντας την εκπαίδευση σε ανομοιογενείς ετικέτες. Επιπλέον, είναι ανθεκτική (robust), καθώς αποκόπτει (trims) το 20% των ακραίων τιμών (outliers) σε κάθε εικόνα, μειώνοντας την ευαισθησία του μοντέλου σε μη ακριβή ετικέτες του ground truth.

Τέλος, η επιτυχία του MiDaS εξαρτάται και από τις τεχνικές **Transfer Learning** που εφαρμόστηκαν. Χρησιμοποιήθηκαν encoders υψηλής χωρητικότητας (π.χ., ResNeXt-101-WSL ή Vision Transformers) που είχαν προ-εκπαιδευτεί σε τεράστια σύνολα δεδομένων (π.χ., ImageNet ή Weakly-Supervised Data) πριν από την εκπαίδευση με στόχο την εκτίμηση του βάθους. Αυτή η προ-εκπαίδευση παρέχει στον κωδικοποιητή εξαιρετικά γενικεύσιμες

αναπαραστάσεις των χαρακτηριστικών της εικόνας, οι οποίες μεταφέρονται στην εργασία εκτίμησης βάθους, ενισχύοντας την ικανότητα του μοντέλου να ερμηνεύει άγνωστες σκηνές.

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.

[4] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler και V. Koltun, *Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer*, 2020. arXiv: 1907.01341 [cs.CV]. διεύθν.: https://arxiv.org/abs/1907.01341.

[5] R. Ranftl, A. Bochkovskiy και V. Koltun, *Vision Transformers for Dense Prediction*, 2021. arXiv: 2103.13413 [cs.CV]. διεύθν.: https://arxiv.org/abs/2103.13413.