Benchmarking Monocular Depth Estimation Models

Νικόλας Σπυρόπουλος
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα, Ελλάδα
el21202@mail.ntua.gr

***Περίληψη*---Η εκτίμηση βάθους από μία εικόνα (Monocular Depth Estimation) αποτελεί θεμελιώδες πρόβλημα στην όραση υπολογιστή, με κρίσιμες εφαρμογές στην αυτόνομη πλοήγηση και την τρισδιάστατη ανακατασκευή. Η παρούσα εργασία εστιάζει στην πρόκληση της γενίκευσης τομέα (domain generalization), εξετάζοντας την εξέλιξη τεσσάρων σύγχρονων μοντέλων: MiDaS, ZoeDepth, Depth-Anything-V2 και Marigold. Αναλύονται οι αρχιτεκτονικές και οι στρατηγικές εκπαίδευσης που επιτρέπουν σε αυτά τα δίκτυα να αποδίδουν σε άγνωστες σκηνές, από τη μίξη ετερογενών δεδομένων του MiDaS και τον συνδυασμό σχετικού-μετρικού βάθους του ZoeDepth, έως τη χρήση συνθετικών δεδομένων στο Depth-Anything-V2 και την αξιοποίηση μοντέλων διάχυσης (diffusion models) στο Marigold. Το κείμενο αυτό αποτελεί το πρώτο μέρος της μελέτης, θέτοντας το θεωρητικό υπόβαθρο για την επακόλουθη πειραματική αξιολόγηση και σύγκριση της απόδοσης των μοντέλων σε πρότυπα σύνολα δεδομένων (benchmarking).**

***Λέξεις Κλειδιά*---Monocular Depth Estimation, Benchmarking, Deep Learning, Computer Vision, Transformers**

I. Εισαγωγη

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.

II. Εκτιμηση Βαθουσ (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 μησαφώς 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.

Α΄. 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, συνθήκες φωτισμού ή διαφορετικούς τύπους σκηνών(πχ εσωτερικές ή εξωτερικές σκηνές). 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.

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

Β΄. 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.

III. Γενικευση Τομεα (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.

IV. 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 διαφορετικού τύπου (inconsistent). 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.

Α΄. Αρχιτεκτονική

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

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)**.

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 τμήματα πληροφορίας) από τα διάφορα στάδια του Transformer. Στην συνέχεια "συναρμολογεί" εκ νέου (**reassemble**) αυτά τα

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

tokens σε αναπαραστάσεις εικόνας πολλαπλών αναλύσεων, τις οποίες επεξεργάζεται για να παραχθεί ο τελικός πυκνός χάρτης βάθους.

Β΄. Εκπαίδευση

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. Επιπλεόν, 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.

V. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth

Το **ZoeDepth** αντιπροσωπεύει εξέλιξη της έρευνας του μοντέλου MiDaS. Ενώ τα MiDaS v1-v3 ήταν εξαιρετικά στην πρόβλεψη του σχετικού βάθους (relative depth) – δηλαδή τη σωστή διάταξη των αντικειμένων στο χώρο– , απέτυχαν να διατηρήσουν τη μετρική κλίμακα (metric scale) (πχ το πραγματικό βάθος σε μέτρα). Το ZoeDepth είναι η πρώτη προσέγγιση που συνδυάζει την γενίκευση του MiDaS με την ικανότητα διατήρησης της μετρικής κλίμακας [6].

Α΄. Αρχιτεκτονική

Το ZoeDepth διατηρεί την αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή, κληρονομώντας τις βέλτιστες πρακτικές του DPT. Προκειμένου να αντιμετωπίσει το μέχρι τότε δυσεπίλυτο πρόβλημα της απόλυτης εκτίμησης βάθους, εισάγει μια κρίσιμη καινοτομία στην κεφαλή πρόβλεψης, το **Metric Bins Module**.

- **Κωδικοποιητής**: Όπως ακριβώς και το MiDaS v3, το ZoeDepth χρησιμοποιεί ViT ως backbone και συγκεκριμένα το **Beit-L**. Έτσι, όπως έχει αποδειχθεί, εξασφαλίζει την καλύτερη δυνατή γενίκευση και μοντελοποίηση των ολικών εξαρτήσεων στην εικόνα.
- **Αποκωδικοποιητής**: Ο αποκωδικοποιητής παραμένει παρόμοιος με αυτόν του DPT, αλλά με μια σημαντική καινοτομία - το **Metric Bins Module (MBM)**. Αντί να προβλέπει το βάθος ως μία συνεχή τιμή (regression), το ZoeDepth προβλέπει πιθανότητες βάθους σε ένα σύνολο **διακριτών bins**, τα οποία αναπαριστούν ένα συγκεκριμένο εύρος βάθους (πχ 0-80m για εξωτερικές σκηνές ή 0-10m για εσωτερικές). Το MBM επιτρέπει την προσαρμογή της θέσης αυτών των bins κατά τη διάρκεια της εκπαίδευσης με μετρικό βάθος (**metric depth fine-tuning**).
- **Router**: Επειδή το μοντέλο έχει γίνει fine-tune σε διαφορετικά datasets για εσωτερικές και εξωτερικές σκηνές και υπάρχουν επομένως, διαφορετικές κεφαλές (**Metric Heads**) για κάθε τύπο σκηνής. Έτσι, είναι αναγκαίο να υπάρχει ένας μηχανισμός που να επιλέγει την κατάλληλη κεφαλή ανάλογα με την είσοδο. Για αυτό το σκοπό, το ZoeDepth εισάγει έναν **Router**, ο οποίος αποτελείται από ένα **Multi Layer Perceptron (MLP)** ταξινομητή που εκπαιδεύεται ταυτόχρονα με το υπόλοιπο μοντέλο.

Β΄. Εκπαίδευση

Η εκπαίδευση του ZoeDepth αναλύεται σε δύο φάσεις. Στην πρώτη φάση γίνεται ένα pre-train εμπνευσμένο από το MiDaS, ενώ στην δεύτερη φάση γίνεται το fine-tune που επιλύει το πρόβλημα της απόλυτης/μετρικής εκτίμησης βάθους.

1) Φάση 1: Προ-εκπαίδευση Σχετικού Βάθους (Relative Depth Pre-training): Η πρώτη φάση ακολουθεί την μεθοδολογία του MiDaS. Εκπαιδεύεται σε 12 datasets με ετικέτες **σχετικού βάθους** χρησιμοποιώντας την ίδια συνάρτηση

απώλειας **Scale- and Shift-Invariant Loss**. Έτσι, χτίζεται ένας ισχυρά γενικεύσιμος κωδικοποιητής, ικανός να ερμηνεύει ποικίλα datasets.

2) Φάση 2: Fine-tuning Μετρικού Βάθους (Metric Depth Fine-tuning): Η δεύτερη φάση εκπαίδευσης είναι μια διαδικασία fine-tuning, με στόχο την ευθυγράμμιση του ήδη γενικεύσιμου κωδικοποιητή με την πραγματική μετρική κλίμακα. Κατά τη διάρκεια αυτής της φάσης, το μοντέλο χρησιμοποιεί δύο υψηλής ποιότητας datasets που περιέχουν μετρικό ground truth -το **KITTI** για εξωτερικές σκηνές και το **NYU Depth V2** για εσωτερικές.

Ουσιαστικά, η εκπαίδευση επικεντρώνεται στο Metric Bins Module (MBM), την νέα κεφαλή του μοντέλου. Ενώ στην πρώτη φάση το μοντέλο έμαθε να τοποθετεί τα αντικείμενα σωστά σε σχετικό βάθος, στη δεύτερη φάση το MBM ενεργοποιείται και βελτιστοποιείται. Το μοντέλο μαθαίνει να προσαρμόζει τις εκπαιδεύσιμες παραμέτρους που ορίζουν τα όρια των διακριτών metric bins, ώστε να αντιστοιχούν ακριβέστερα στις πραγματικές μετρικές τιμές.

Χρησιμοποιώντας μια τυπική μετρική συνάρτηση απώλειας (metric loss), το μοντέλο μαθαίνει την σωστή κλίμακα, επιτυγχάνοντας πλέον ακριβή πρόβλεψη βάθους σε μέτρα. Το αποτέλεσμα είναι ένα μοντέλο που διατηρεί την εξαιρετική ικανότητα zero-shot transfer που αποκτήθηκε από την προ-εκπαίδευση, ενώ ταυτόχρονα μπορεί να δίνει αξιόπιστες μετρικές τιμές σε διαφορετικού είδους εικόνες εισόδου.

VI. Depth-Anything-V2

Το Depth Anything V2 (DA-V2) αποτελεί ένα σημαντικό βήμα προς την επίτευξη πιο **λεπτών (finer-grained)** και πιο **ανθεκτικών (robust)** προβλέψεων σχετικού βάθους. Η καινοτομία του δεν έγκειται σε περίπλοκες τεχνικές, αλλά σε μερικές πρακτικές που αφορούν την κλιμάκωση των δεδομένων και τη διαδικασία εκπαίδευσης [7] [8].

Α΄. Αρχιτεκτονική

Η αρχιτεκτονική του DA-V2 ακολουθεί το πρότυπο Κωδικοποιητή-Αποκωδικοποιητή που καθιερώθηκε από τα DPT/MiDaS, δίνοντας έμφαση στη χρήση ενός Vision Transformer για τη μέγιστη ικανότητα γενίκευσης.

- **Κωδικοποιητής**: Χρησιμοποιείται ένας μεγάλος Vision Transformer (ViT), ο **ViT-Large** που περιέχει 335M παραμέτρους. Είναι προ-εκπαιδευμένος μέσω ενός σχήματος **αυτο-επιβλεπόμενης μάθησης (Self Supervised Learning)** σύμφωνα με την μέθοδο που ορίζει το **DINOv2** [9]. Η συγκεκριμένη μέθοδος επιτρέπει στους Transformers να μαθαίνουν ισχυρές, υψηλής ποιότητας αναπαραστάσεις χωρίς ανθρώπινη επισήμανση (labels), καθιστώντας τον κωδικοποιητή εξαιρετικά αποτελεσματικό στην εξαγωγή πλούσιων γεωμετρικών χαρακτηριστικών που απαιτούνται για την εκτίμηση βάθους. Τέλος, μέσω του σχήματος Teacher-Student, αφού δημιουργηθεί το μεγάλο μοντέλο **Teacher** με τον ViT-Large, η έρευνα δίνει μεγάλη έμφαση στο πώς μπορεί να μεταφερθεί αυτή η γνώση (**Knowledge Distillation**) σε μικρότερα **Student** μοντέλα, πχ ViT-Small και ViT-Base, εμφανώς λιγότερων παραμέτρων.
- **Αποκωδικοποιητής**: Η δομή του αποκωδικοποιητή είναι παρόμοια με αυτή του DPT. Είναι υπεύθυνος για τη σύντηξη (fusion) των χαρακτηριστικών πολλαπλών επιπέδων (multi-scale features) που εξάγονται από τον κωδικοποιητή ViT. Χρησιμοποιεί **ανοδική δειγματοληψία (upsampling)** και συνελικτικές μονάδες για να ανακτήσει τις χωρικές λεπτομέρειες και να παράξει τον τελικό πυκνό χάρτη βάθους, διατηρώντας την ανάλυση και τις λεπτές ακμές των αντικειμένων.

Β΄. Εκπαίδευση

Η εκπαίδευση του Depth-Anything-V2 διαφέρει ριζικά από τα προηγούμενα μοντέλα, καθώς απορρίπτει όλα τα datasets με πραγματικές μετρικές ετικέτες (labeled real images) και βασίζεται σε μια διαδικασία τριών βημάτων που εστιάζει στα **συνθετικά δεδομένα** και στη **Μετάδοση Γνώσης**.

1) Βήμα 1: Αποκλειστική χρήση Συνθετικών Δεδομένων (Synthetic Data Training): Το DA-V2 αντικατέστησε όλες τις επισημασμένες πραγματικές εικόνες με συνθετικές εικόνες. Το **Teacher Model** εκπαιδεύτηκε αποκλειστικά σε ένα μαζικό σύνολο συνθετικών δεδομένων (BlendedMVS, TartanAir, HRWSI, κ.ά.), τα οποία παρέχουν τέλειο και καθαρό ground truth βάθους, χωρίς τα σφάλματα και τους θορύβους των πραγματικών αισθητήρων (πχ LiDAR). Αυτή η στρατηγική επιτρέπει στο μοντέλο να μάθει τις βασικές γεωμετρικές αρχές με μεγάλη ακρίβεια.

2) Βήμα 2: Παραγωγή Ψευδο-ετικετών (Pseudo-Labeling): Το μοντέλο που εκπαιδεύτηκε στα συνθετικά δεδομένα (Teacher Model) χρησιμοποιείται για να **παράγει ψευδο-ετικέτες (pseudo-labels)** σε μεγάλης κλίμακας πραγματικές εικόνες που δεν έχουν ετικέτες (unlabeled real images). Με αυτόν τον τρόπο, το μοντέλο γεφυρώνει το χάσμα μεταξύ των συνθετικών και των πραγματικών δεδομένων. Οι ψευδο-ετικέτες διατηρούν την υψηλή ακρίβεια των γεωμετρικών κανόνων που μάθανε από τα συνθετικά, αλλά εφαρμόζονται στην πραγματική κατανομή εικόνων του κόσμου (real-world distribution).

3) Βήμα 3: Μετάδοση Γνώσης: Η Μετάδοση Γνώσης είναι η διαδικασία με την οποία η γνώση που αποκτήθηκε από το μεγάλο μοντέλο Teacher μεταφέρεται σε ένα μικρότερο και πιο αποδοτικό μοντέλο Student. Ουσιαστικά, το Teacher Model, το οποίο έχει ήδη μάθει από συνθετικά δεδομένα και έχει δημιουργήσει ψευδο-ετικέτες σε μαζικές, μη επισημασμένες πραγματικές εικόνες, χρησιμοποιείται για να "διδάξει" το Student Model (πχ, ViT-Small, 25M παραμέτρων). Το Student εκπαιδεύεται με μια πιο "χαλαρή" συνάρτηση απώλειας ώστε να μιμηθεί την πρόβλεψη βάθους του Teacher. Αυτή η μέθοδος επιτρέπει στο Student να διατηρεί την υψηλή ακρίβεια, την ανθεκτικότητα και την ικανότητα παραγωγής λεπτών λεπτομερειών του Teacher, επιτυγχάνοντας παράλληλα πολύ **ταχύτερη εξαγωγή (inference speed)** και σημαντικά **μειωμένο αριθμό παραμέτρων**. Με αυτόν

τον τρόπο, το Depth-Anything-V2 μπορεί να προσφέρει μοντέλα κατάλληλα για **εφαρμογές πραγματικού χρόνου**.

VII. Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

Το **Marigold** σηματοδοτεί μια κομβική και ριζική αλλαγή στην προσέγγιση της εκτίμησης βάθους, απομακρύνοντας την έρευνα από τα παραδοσιακά ViT/CNN μοντέλα (MiDaS, ZoeDepth, Depth Anything) που βασίζονται σε encoders οπτικών χαρακτηριστικών. Η καινοτομία του Marigold έγκειται στην αξιοποίηση της εκτεταμένης οπτικής γνώσης για τον φυσικό κόσμο που έχει αποκτηθεί από ένα προ-εκπαιδευμένο μοντέλο διάχυσης (**pre-trained Diffusion Model**), όπως το **Stable Diffusion**. Έτσι χρησιμοποιεί παίρνει αυτό το ισχυρό generative μοντέλο εικόνας και το κάνει **fine-tune για την εργασία της εκτίμησης βάθους** [10].

Α΄. Αρχιτεκτονική

Η αρχιτεκτονική του Marigold είναι θεμελιωδώς διαφορετική από τα προηγούμενα μοντέλα, καθώς βασίζεται στην αξιοποίηση ενός προ-εκπαιδευμένου **Μοντέλου Διάχυσης (Diffusion Model)**, του Stable Diffusion, αντί να εκπαιδευτεί από την αρχή ένα Vision Transformer. Συγκεκριμένα, βασίζεται στο **Noise Predictor** του Stable Diffusion, το οποίο είναι ένα U-Net δίκτυο.

- Η ακρίβεια του Marigold δεν προέρχεται από την εκπαίδευση σε δεδομένα βάθους, αλλά από την πλούσια οπτική κατανόηση της δομής και της σύνθεσης των σκηνών που έχει αποκτήσει το U-Net από την προεκπαίδευσή του σε **"internet-scale"** πλήθος εικόνων (πχ στο σύνολο δεδομένων LAION-5B).
- **Ρόλος του U-Net**: Η αρχιτεκτονική του U-Net είναι σχεδιασμένη για πυκνές χωρικές εξαρτήσεις, καθιστώντας το κατάλληλο για την εργασία της εκτίμησης βάθους, ακόμα και αν ο αρχικός του ρόλος ήταν η πρόβλεψη θορύβου.
- **Είσοδος**: Το δίκτυο τροποποιείται ώστε να δέχεται την **RGB εικόνα** της σκηνής, μαζί με τον προστιθέμενο θόρυβο και την Τιμή Χρόνου (t) από τη διαδικασία διάχυσης.
- **Έξοδος**: Η τελική έξοδος του U-Net είναι ένας **χάρτης δυσαναλογίας (disparity map)**, ο οποίος στη συνέχεια επεξεργάζεται για να δώσει το τελικό βάθος.

Β΄. Εκπαίδευση

Η εκπαίδευση του Marigold εστιάζει στο fine-tuning της ήδη υπάρχουσας γνώσης, ενώ η πρόβλεψη αξιοποιεί τον επαναληπτικό μηχανισμό της διάχυσης.

- **Fine-tuning σε συνθετικά δεδομένα**: Το μοντέλο δεν εκπαιδεύεται σε μαζικά, ετερογενή datasets βάθους (όπως το MiDaS), αλλά χρησιμοποιεί έναν σχετικά μικρό αριθμό συνθετικών δεδομένων, περίπου 74K εικόνες, ώστε να "μάθει" να παράγει ακριβείς χάρτες βάθους. Ο σκοπός είναι η "μετάφραση" της ήδη κωδικοποιημένης οπτικής γνώσης (από το Stable Diffusion) στον κανόνα αντιστοίχισης μεταξύ RGB εικόνας και βάθους. Αυτό το πετυχαίνει χρησιμοποιώντας απλές τεχνικές κανονικοποίησης (regularization), όπως L1/L2 απώλειες.
- **Επαναληπτική Σύνθεση (Iterative Ensembling)**: Κατά την φάση της **πρόβλεψης (inference)**, το Marigold διατηρεί τον επαναληπτικό χαρακτήρα του diffusion model και αντί για μόνο ένα πέρασμα, εκτελείται για 10 **βήματα δειγματοληψίας (sampling steps)**. Έτσι, ο τελικός χάρτης βάθους προκύπτει από την **σύνθεση (ensembling)** όλων των επαναληπτικών προβλέψεων, με αποτέλεσμα την μεγάλη βελτίωση της ποιότητας και της ανθεκτικότητας (robustness) της πρόβλεψης.

Αναφορες

[1] J. Zhang, ``Survey on monocular metric depth estimation,'' 2025. [Online]. Available: https://arxiv.org/abs/2501.11841

[2] D. Eigen, C. Puhrsch, and R. Fergus, ``Depth map prediction from a single image using a multi-scale deep network,'' 2014. [Online]. Available: https://arxiv.org/abs/1406.2283

[3] Y. Yao, G. Wu, K. Jiang, S. Liu, J. Kuai, X. Liu, and J. Jiang, *Improving Domain Generalization in Self-supervised Monocular Depth Estimation via Stabilized Adversarial Training*. Springer Nature Switzerland, Nov. 2024, p. 183–201. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-72691-0_11

[4] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, ``Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,'' 2020. [Online]. Available: https://arxiv.org/abs/1907.01341

[5] R. Ranftl, A. Bochkovskiy, and V. Koltun, ``Vision transformers for dense prediction,'' 2021. [Online]. Available: https://arxiv.org/abs/2103.13413

[6] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, ``Zoedepth: Zero-shot transfer by combining relative and metric depth,'' 2023. [Online]. Available: https://arxiv.org/abs/2302.12288

[7] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, ``Depth anything: Unleashing the power of large-scale unlabeled data,'' in *CVPR*, 2024.

[8] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, ``Depth anything v2,'' *arXiv:2406.09414*, 2024.

[9] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, ``Dinov2: Learning robust visual features without supervision,'' 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[10] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, ``Repurposing diffusion-based image generators for monocular depth estimation,'' 2024. [Online]. Available: https://arxiv.org/abs/2312.02145