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. Αντικείμενο της εργασίας αποτελεί η 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. Στην συνέχεια, πραγματοποιείται πειραματική αξιολόγηση (benchmarking) των μοντέλων σε συνθήκες μηδενικής εκμάθησης (zero-shot). Για την αξιολόγηση χρησιμοποιήθηκαν τρία ετερογενή σύνολα δεδομένων: το NYU Depth V2 (εσωτερικοί χώροι), το KITTI (εξωτερικοί χώροι/αυτόνομη οδήγηση) και το iBims-1 (για έλεγχο ακρίβειας σε ακμές και διαφανείς επιφάνειες). Η μεθοδολογία περιλαμβάνει την ευθυγράμμιση προβλέψεων (scale alignment) για μοντέλα σχετικού βάθους και τη χρήση μετρικών όπως AbsRel, δ_1 και SILog. Τα αποτελέσματα δείχνουν την υπεροχή των θεμελιωδών μοντέλων (foundation models) όπως το Depth-Anything-V2 στη γενίκευση, ενώ αναδεικνύουν τους περιορισμούς παλαιότερων προσεγγίσεων.**

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

I. Εισαγωγη

Η κατανόηση της τρισδιάστατης δομής ενός σκηνικού αποτελεί τον κύριο 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 σημαντική 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

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

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.

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 Transformer (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

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

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 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. 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.

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" συνάρ-

τηση 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 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.

VIII. Αναλυτικη Περιγραφη Μεθοδολογιας

Έχοντας πλέον καλύτερη γνώση τόσο των αρχιτεκτονικών όσο και των τρόπων εκπαίδευσης των μοντέλων, πριν παρουσιαστεί η πειραματική μέθοδος για την σύγκριση και αξιολόγησή τους, είναι αναγκαίο να γίνει σαφής η επιλογή των συνόλων δεδομένων (**datasets**) που χρησιμοποιήθηκαν. Η αξιολόηση αυτή έγινε σε συνθήκες **zero-shot**, δηλαδή δεν προηγήθηκε καμία διαδικασία fine-tuning ή επεξεργασία των αποτελεσμάτων, ώστε να αναδειχθούν οι ικανότητες ή οι περιορισμοί των μοντέλων σαν να χρησιμοποιούνταν για οποιαδήποτε εικόνα σε οποιοδήποτε πλαίσιο (**in-the-wild**). Χρησιμοποιήθηκαν εικόνες από τρία διαφορετικά σύνολα δεδομένων[3] [4] [5] ώστε να φανεί η ικανότητα κάθε μοντέλου στην γενίκευση τομέα.

Α΄. Επιλογή Συνόλων Δεδομένων

Για την κάλυψη ενός μεγάλου φάσματος σκηνών και την αξιολόγηση της ευρωστίας (robustness) των μοντέλων, επιλέχθηκαν τρία ετερογενή σύνολα δεδομένων (datasets). Λόγω του μεγάλου αριθμού δεδομένων που περιείχαν τα τελυεταία δύο σύνολα και του περιορισμένου αριθμού πόρων, επιλέχθηκαν τυχαία 100 δείγματα (sampling) από κάθε ένα και έτσι κάθε σύνολο δεδομένων για την αξιολόγηση περιέχει 100 εικόνες.

Β΄. IBims-1

Το σύνολο δεδομένων **IBims-1** [11] επιλέχθηκε ως ένα εξειδικευμένο εργαλείο αξιολόγησης για σκηνές εσωτερικού χώρου, εστιάζοντας στην ποιότητα των λεπτομερειών και την ορθότητα των ακμών. Σε αντίθεση με τα τυπικά datasets εκπαίδευσης, το IBims-1 περιέχει σκηνές με με διαφανείς επιφάνειες και έντονες μεταβάσεις βάθους, επιτρέποντας τον έλεγχο της συμπεριφοράς των μοντέλων σε

[3]https://www.kaggle.com/datasets/patiencechewyeecheah/ibims-1

[4]https://www.kaggle.com/datasets/soumikrakshit/nyu-depth-v2

[5]https://www.kaggle.com/datasets/artemmmtry/kitti-depth-prediction-evaluation

δύσκολες συνθήκες. Για την παρούσα εργασία χρησιμοποιήθηκε το υποσύνολο "core" (περιέχει 100 σκηνές), χωρίς περαιτέρω χωρική περικοπή (cropping), προκειμένου να εξεταστεί η ικανότητα των αλγορίθμων να διαχειρίζονται ολόκληρη την πληροφορία της σκηνής και να διατηρούν τη δομική συνοχή σε αντικείμενα με περίπλοκα όρια ακόμα και στις άκρες της κάθε εικόνας.

Γ΄. NYU Depth V2

Το σύνολο δεδομένων **NYU Depth V2** [12] αποτελεί ένα καθιερωμένο πρότυπο αναφοράς για την εκτίμηση βάθους σε εσωτερικούς χώρους, περιλαμβάνοντας πυκνούς χάρτες βάθους (dense depth maps) και εικόνες RGB από ποικίλες οικιακές και επαγγελματικές σκηνές ληφθείσες με αισθητήρα Microsoft Kinect. Ο κύριος σκοπός της χρήσης του είναι η συγκριτική αξιολόγηση της γενίκευσης των μοντέλων σε τυπικά περιβάλλοντα εσωτερικού χώρου. Στο πλαίσιο της πειραματικής διαδικασίας, εφαρμόστηκε κατά την αξιολόγηση η τεχνική Eigen Crop, κάνοντας χρήση της μάσκας εγκυρότητας που παρέχεται με το dataset. Η επεξεργασία αυτή είναι απαραίτητη για να εξαιρεθούν από τον υπολογισμό του σφάλματος τα pixels στα όρια της εικόνας όπου ο αισθητήρας δεν παρέχει έγκυρες μετρήσεις ή υπάρχουν παραμορφώσεις προβολής.

Δ΄. KITTI

Για την αξιολόγηση σε εξωτερικούς χώρους χρησιμοποιήθηκε το σύνολο δεδομένων **KITTI** [13], το οποίο παρέχει αραιούς χάρτες βάθους (sparse depth maps) προερχόμενους από αισθητήρες LiDAR Velodyne. Σκοπός της χρήσης του είναι η εξέταση της απόδοσης των μοντέλων σε μεγάλες αποστάσεις και σε δυναμικά περιβάλλοντα με έντονες φωτιστικές αλλαγές. Λόγω της διάταξης των αισθητήρων στο όχημα καταγραφής, οι μετρήσεις στο πάνω μέρος (ουρανός) και στα πλάγια της εικόνας είναι συχνά αναξιόπιστες. Για τον λόγο αυτό, εφαρμόστηκε η τυπική διαδικασία περικοπής Garg Crop, η οποία περιορίζει την περιοχή αξιολόγησης στο κεντρικό τμήμα της εικόνας όπου τα δεδομένα ground truth είναι πυκνά και αξιόπιστα, διασφαλίζοντας έτσι μια δίκαιη σύγκριση για τα μοντέλα που δεν έχουν εκπαιδευτεί ειδικά σε αυτό το πεδίο.

Ε΄. Προεπεξεργασία και Ευθυγράμμιση

Μια σημαντική πρόκληση στη σύγκριση διαφορετικών μοντέλων είναι ότι ορισμένα (π.χ. MiDaS) παράγουν σχετικό βάθος (relative depth) χωρίς φυσικές μονάδες, ενώ άλλα (π.χ. ZoeDepth) παράγουν μετρικό βάθος (metric depth). Για να γίνει δυνατή και δίκαιη η σύγκριση μεταξύ των δύο κατηγοριών μοντέλων ακολουθήθηκε η εξής διαδικασία:

- Για μοντέλα **σχετικού βάθους**: Εφαρμόζεται ευθυγράμμιση κλίμακας και μετατόπισης (Scale and Shift Alignment) με τη μέθοδο των Ελαχίστων Τετραγώνων (Least Squares), ώστε η πρόβλεψη να προσαρμοστεί στο εύρος τιμών του ground truth:

$$D_{aligned} = sD_{pred} + t$$

όπου s και t είναι οι παράμετροι κλίμακας και μετατόπισης.
- Για μοντέλα **μετρικού βάθους**: Οι προβλέψεις αξιολογούνται απευθείας, χωρίς καμία προσαρμογή, καθώς στόχος είναι να ελεγχθεί η ικανότητά τους να εκτιμούν τις πραγματικές αποστάσεις.
- Χειρισμός **Disparity**: Για μοντέλα που εξάγουν disparity (αντιστρόφως ανάλογο του βάθους) αντί για depth map, εφαρμόζεται αντιστροφή $(1/x)$ πριν την αξιολόγηση.

Ϛ΄. Μετρικές Αξιολόγησης

Για την ποσοτική αποτίμηση των αποτελεσμάτων χρησιμοποιήθηκαν τρεις καθιερωμένες μετρικές στη βιβλιογραφία της εκτίμησης βάθους. Για όλες τις παρακάτω μετρικές έστω d_i η τιμή βάθους του ground truth pixel i, $\hat{d}_i$ η προβλεπόμενη τιμή από το μοντέλο και N ο συνολικός αριθμός των έγκυρων pixels (έγκυρα pixels θεωρούνται όσα δεν έχουν αποβληθεί μέσω των μασκών (masks) που δίνονται από κάθε σύνολο δεδομένων).

- Απόλυτο Σχετικό Σφάλμα (Absolute Relative Error - AbsRel): δίνεται από τον τύπο

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - \hat{d}_i|}{d_i}$$

  Το **AbsRel** εκφράζει το σφάλμα ως ποσοστό της πραγματικής απόστασης. Είναι σημαντική μετρική γιατί για παράδειγμα ένα σφάλμα 50 εκατοστών είναι αμελητέο αν το αντικείμενο βρίσκεται στα 50 μέτρα, αλλά καταστροφικό αν το αντικείμενο βρίσκεται στο 1 μέτρο. Έτσι, κανονικοποιεί το σφάλμα ώστε να έχει την ίδια βαρύτητα ανεξάρτητα από την απόσταση. Όσο πιο κοντά είναι η τιμή του στο 0, τόσο καλύτερη είναι η επίδοση του μοντέλου.
- Ακρίβεια Δέλτα: δίνεται από τον τύπο

$$\delta_1 = \frac{1}{N}\sum_{i=1}^{N} 1(max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}) < 1.25)$$

  Το δ_1 αναδεικνύει το ποσοστό των "επιτυχημένων" pixels. Ένα pixel θεωρείται σωστό ή επιτυχημένο αν η πρόβλεψη δεν απέχει περισσότερο από 25% από την πραγματική τιμή. Είναι ένας δείκτης της αξιοπιστίας (robustness) του μοντέλου. Δεν δίνει βάρος στο πόσο μεγάλο είναι το λάθος στα αποτυχημένα pixels (πχ outliers), αλλά πόσο συχνά το μοντέλο "πέφτει μέσα". Επομένως, η τιμή του πρέπει να είναι όσο πιο κοντά στο 1 γίνεται για να θεωρείται καλό το μοντέλο βάσει αυτής της μετρικής.
- Λογαριθμικό Σφάλμα Ανεξαρτήτου Κλίμακας (Scale Invariant Logarithmic Error - SILog): Έστω $\Delta_i = log\delta_i - log\hat{\delta}_i$ και ο συντελεστής βάρους $\lambda = 1$. Δίνεται από τον τύπο

$$SILog = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Delta_i^2 - \frac{\lambda}{N^2}(\sum_{i=1}^{N}\Delta_i)^2} \times 100$$

Το **SILog** μετράει τη γεωμετρική συνέπεια της σκηνής, αγνοώντας το απόλυτο μέγεθος. Πολλά μοντέλα (ειδικά τα relative depth models) μπορεί να καταλάβουν σωστά ότι ένα αντικείμενο βρίσκεται μπροστά από ένα άλλο, αλλά να κάνουν λάθος στο μέγεθος ολόκληρης της σκηνής. Έτσι, το SILog "τιμωρεί" τα λάθη στις σχέσεις των αντικειμένων (δομικά λάθη), αλλά "συγχωρεί" το μοντέλο αν έχει κάνει λάθος στην γενική κλίμακα (scale) όλης της εικόνας. Όσο πιο κοντά είναι η τιμή του στο 0, τόσο καλύτερα έχει κατανοήσει το μοντέλο την τριδιάστατη δομή της σκηνής.

Ζ΄. Περιγραφή πειραματικής διαδικασίας

Η πειραματική διαδικασία σχεδιάστηκε ως ένα ενιαίο pipeline αξιολόγησης για να διασφαλιστεί η δίκαιη και συνεπής σύγκριση όλων των μοντέλων. Η ανάπτυξη και εκτέλεση των πειραμάτων πραγματοποιήθηκε η πλατφόρμα που παρέχει το Kaggle σε περιβάλλον Python με χρήση των βιβλιοθηκών transformers και diffusers. Από τις επιλογές σε hardware που προσφέρει το Kaggle ελεύθερα, αξιοποιήθηκε η κάρτα γραφικών (GPU) NVIDIA Tesla P100 (16GB VRAM), στοιχείο απαραίτητο για την αποδοτική εκτέλεση (inference) των βαρύτερων μοντέλων όπως το Marigold.

Όλα τα checkpoints για τα μοντέλα που χρησιμοποιήθηκαν είναι από το Hugging Face και πιο συγκεκριμένα:

- **DAV2-Large**[6] (relative depth): Είναι το βασικό μοντέλο Depth Anything V2 όπως παρουσιάστηκε και στο θεωρητικό κομμάτι. Είναι το μεγαλύτερο από τα διαθέσιμα μοντέλα ώστε να αξιολογηθεί το "καλύτερο" από τα όλα τα διαθέσιμα.
- **DAV2-Indoor-Metric**[7] (metric depth): Το backbone του είναι το Depth Anything V2 αλλά έχει γίνει fine-tune σε μετρικά δεδομένα εσωτερικών χώρων.
- **DAV2-Outdoor-Metric**[8] (metric depth): Παρόμοια, έχει backbone το κλασσικό Depth Anything V2 και έχει γίνει fine-tune σε μετρικά δεδομένα εξωτερικών χώρων.
- **ZoeDepth**[9] (metric depth): Έχει ως πυρήνα το κλασσικό ZoeDepth από το αρχικό paper και έχει γίνει fine-tune στα σύνολα Nyu και KITTI.
- **MiDaS-3.0**[10] (relative depth): Η έκδοση 3.0 του MiDaS.
- **MiDaS-3.1**[11] (relative depth): Η έκδοση 3.1 του MiDaS με τον Beit transformer ως backbone.
- **Marigold-1.1**[12] (relative depth): Η έκδοση 1.1 του Marigold. Χρησιμοποιήθηκαν μόνο 4 βήματα αποθορυβοποίησης (denoising steps) λόγω των περιορισμένων πόρων.

[6]https://huggingface.co/depth-anything/Depth-Anything-V2-Large-hf

[7]https://huggingface.co/depth-anything/Depth-Anything-V2-Metric-Indoor-Large-hf

[8]https://huggingface.co/depth-anything/Depth-Anything-V2-Metric-Outdoor-Large-hf

[9]https://huggingface.co/Intel/zoedepth-nyu-kitti

[10]https://huggingface.co/Intel/dpt-large

[11]https://huggingface.co/Intel/dpt-beit-large-512

[12]https://huggingface.co/prs-eth/marigold-depth-v1-1

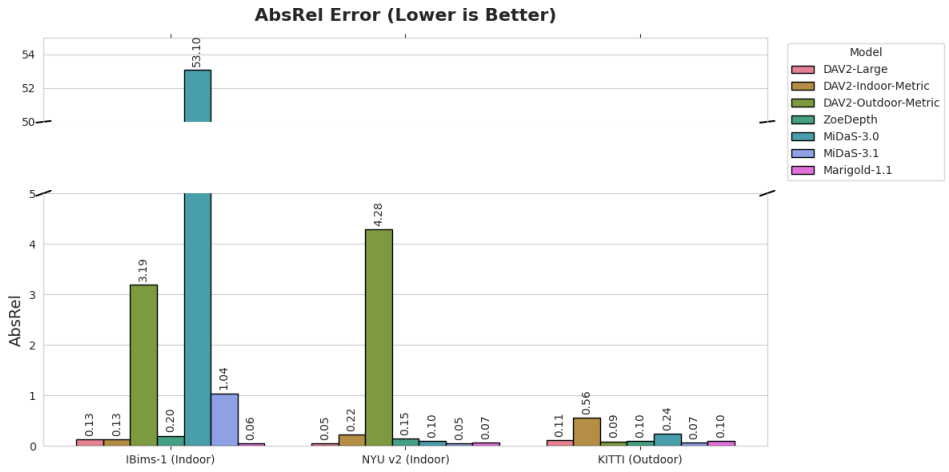
Για κάθε ζεύγος εικόνας εισόδου I και χάρτη βάθους αναφοράς (Ground Truth - D_{gt}) από τα σύνολα δεδομένων δοκιμής, ακολουθήθηκε η εξής διαδικασία:

- Προσαρμογή Εισόδου (Input Resizing): Η εικόνα RGB αναδιατάσσεται (resize) στις διαστάσεις που απαιτεί η αρχιτεκτονική του εκάστοτε μοντέλου (π.χ. 518×518 για το Depth-Anything-V2), διατηρώντας την αναλογία διαστάσεων όπου αυτό είναι εφικτό.
- Εκτέλεση Μοντέλου (Inference): Η εικόνα τροφοδοτείται στο μοντέλο και λαμβάνεται στην έξοδο η πρόβλεψη του μοντέλου είτε ως depth map είτε ως disparity map.
- Ανάκτηση Ανάλυσης (Resolution Restoration): Η παραγόμενη πρόβλεψη βάθους D_{pred} επαναφέρεται στην αρχική ανάλυση της εικόνας εισόδου ($H \times W$) μέσω διγραμμικής παρεμβολής (bilinear interpolation), ώστε να υπάρχει αντιστοίχιση pixel-προς-pixel με το Ground Truth.
- Διαχείριση Τιμών (Value Handling):
  - Εάν το μοντέλο παράγει disparity, οι τιμές αντιστρέφονται ($1/x$).
  - Εάν το μοντέλο παράγει σχετικό βάθος, εφαρμόζεται η ευθυγράμμιση ελαχίστων τετραγώνων (scale and shift alignment) που περιγράφηκε προηγουμένως.
- Εφαρμογή Μάσκας (Masking & Cropping): Δημιουργείται μια δυαδική μάσκα εγκυρότητας M, η οποία προκύπτει από την τομή των έγκυρων pixels του Ground Truth και της περιοχής ενδιαφέροντος που ορίζει το κάθε dataset (Eigen Crop για το NYU, Garg Crop για το KITTI).
- Υπολογισμός Σφαλμάτων: Οι μετρικές αξιολόγησης υπολογίζονται αποκλειστικά για τα pixels που ανήκουν στη μάσκα M, αγνοώντας τις περιοχές χωρίς πληροφορία ή τις περιοχές που έχουν εξαιρεθεί λόγω της διαδικασίας περικοπής.

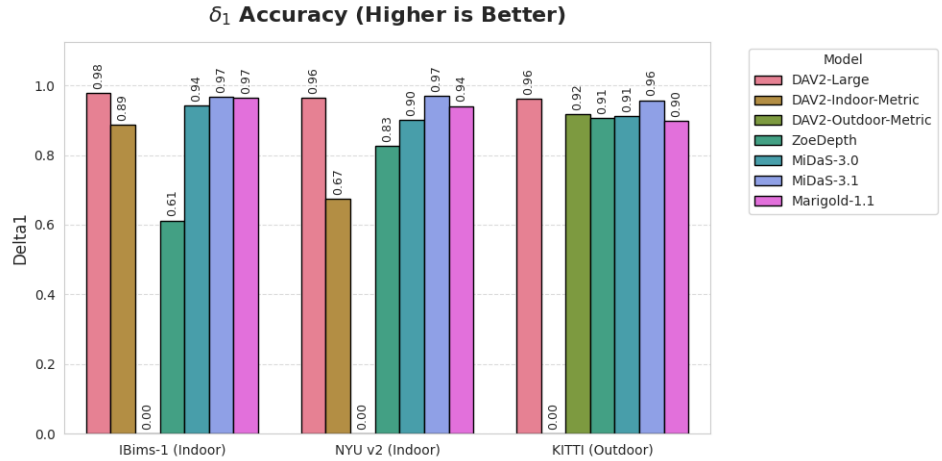
IX. Πειραματικα Αποτελεσματα

Ακολουθώντας την διαδικασία που μόλις περιγράφηκε, υπολογίζονται οι μέσοι όροι για κάθε μετρική από κάθε εικόνα σε κάθε σύνολο δεδομένων.

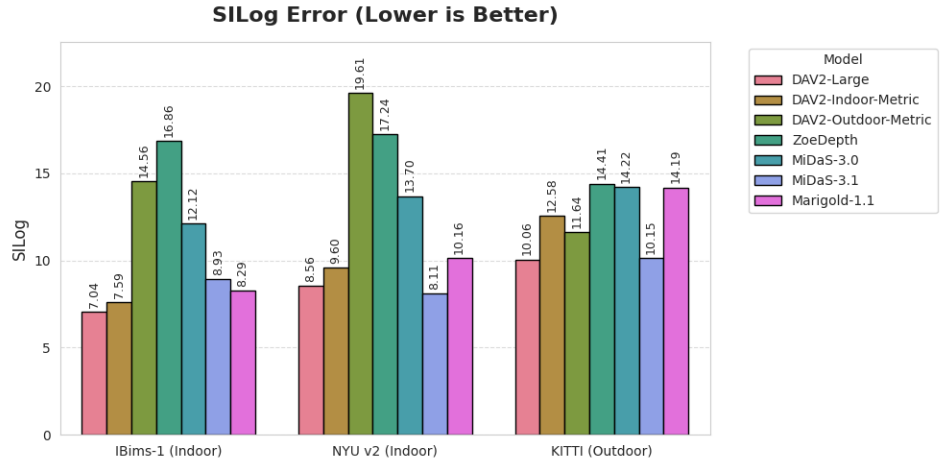
1) Ποσοτική Αξιολόγηση: Στον Πίνακα Ι και στο Σχήμα 1 παρουσιάζεται η συνολική απόδοση των μοντέλων. Από την ανάλυση των μετρήσεων προκύπτουν μερικά άμεσα συμπεράσματα. Πρώτον, το μοντέλο DAV2-Large επιδεικνύει την υψηλότερη ευρωστία (robustness) στο σύνολο των πειραμάτων, επιτυγχάνοντας κορυφαίες τιμές τόσο στα indoor όσο και στα outdoor datasets. Αυτό επιβεβαιώνει την υπεροχή των foundation models που έχουν εκπαιδευτεί σε τεράστια κλίμακα δεδομένων. Δεύτερον, παρατηρείται ένα ενδιαφέρον φαινόμενο απόκλισης μεταξύ των μετρικών AbsRel και δ_1. Συγκεκριμένα, σε ορισμένα πειράματα (όπως φαίνεται στα διαγράμματα), η τιμή του AbsRel εμφανίζεται αυξημένη, υποδηλώνοντας μεγάλο μέσο σφάλμα, παρόλο που η ακρίβεια δ_1 παραμένει σε υψηλά επίπεδα



(α′) Απόλυτο Σχετικό Σφάλμα (AbsRel ↓)



(β′) Ακρίβεια δ_1 (↑)



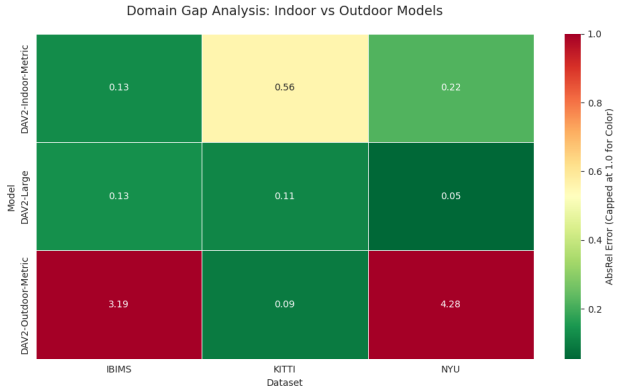
(γ′) Σφάλμα SILog (↓)

Σχήμα 1: Συγκριτική απεικόνιση των μετρικών.

Πίνακας I: Συγκριτικα Αποτελεσματα Αξιολογησης Zero-Shot (Lower is better ↓, Higher is better ↑)

Model	IBims-1 (Indoor/Specialized)			NYU Depth V2 (Indoor)			KITTI (Outdoor)		
	AbsRel ↓	δ_1 ↑	**SILog ↓**	**AbsRel ↓**	δ_1 ↑	**SILog ↓**	**AbsRel ↓**	δ_1 ↑	**SILog ↓**
DAV2-Large	0.135	**0.978**	**7.037**	0.054	0.964	8.557	0.113	**0.962**	**10.06**
DAV2-Indoor-Metric	0.126	0.886	7.589	0.219	0.674	9.598	0.560	0.001	12.57
DAV2-Outdoor-Metric	3.185	0.001	14.55	4.283	0.000	19.60	0.091	0.917	11.63
ZoeDepth	0.200	0.611	16.86	0.146	0.826	17.24	0.104	0.906	14.40
MiDaS-3.0	53.10	0.941	12.12	0.102	0.900	13.69	0.239	0.912	14.21
MiDaS-3.1	1.042	0.966	8.930	**0.052**	**0.969**	**8.111**	**0.065**	0.957	10.15
Marigold-1.1	**0.055**	0.965	8.292	0.071	0.940	10.16	0.103	0.898	14.19

(>0.90). Αυτή η συμπεριφορά οφείλεται στην παρουσία ακραίων τιμών (outliers). Επειδή το AbsRel είναι μέση τιμή, επηρεάζεται έντονα από λίγα pixels με πολύ μεγάλο σφάλμα (π.χ. σε καθρέφτες ή στον ουρανό), ενώ το δ_1, ως ποσοστιαία μετρική, αγνοεί το μέγεθος του σφάλματος εφόσον αυτό ξεπερνά το όριο του 25%. Συνεπώς, η υψηλή τιμή του δ_1 είναι πιο αντιπροσωπευτική της γενικής συμπεριφοράς του μοντέλου στην πλειονότητα των pixels της εικόνας. Τέλος, μπορούμε να δούμε ότι το DAV2-Indoor-Metric αποτυγχάνει στο σύνολο KITTI που περιέχει εξωτερικές σκηνές και αντίστοιχα το DAV2-Outdoor-Metric αποτυγχάνει στα άλλα δύο σύνολα που περιέχουν σκηνές εσωτερικών χώρων.



Σχήμα 2: Οπτικοποίηση του χάσματος πεδίου (Domain Gap). Το διάγραμμα απεικονίζει τη πτώση απόδοσης (αύξηση σφάλματος) των εξειδικευμένων μοντέλων (Metric-Indoor/Outdoor) όταν καλούνται να λειτουργήσουν σε άγνωστο περιβάλλον, συγκριτικά με τη σταθερότητα των μοντέλων σχετικού βάθους (DAV2-Large).

2) Ανάλυση Χάσματος Πεδίου (Domain Gap Analysis): Ένα κρίσιμο ερώτημα της εργασίας αφορά τη δυνατότητα γενίκευσης. Το Σχήμα 2 απεικονίζει τη μεταβολή της απόδοσης κατά την εναλλαγή μεταξύ εσωτερικών και εξωτερικών σκηνών.

Όπως αναμενόταν, τα μοντέλα που εκπαιδεύτηκαν με αποκλειστικό στόχο το μετρικό βάθος σε συγκεκριμένο πεδίο (Specialized Metric Models), αποτυγχάνουν πλήρως όταν καλούνται να γενικεύσουν. Συγκεκριμένα, το DAV2-Metric-Outdoor καταρρέει στο NYU Depth V2, ενώ αντίστοιχα το DAV2-Metric-Indoor αδυνατεί να αποδώσει στο KITTI. Αντιθέτως, τα μοντέλα σχετικού βάθους (Relative Depth), και ειδικά το DAV2-Large και το MiDaS-3.1, γεφυρώνουν επιτυχώς το χάσμα (Domain Gap), καθώς μαθαίνουν δομικά χαρακτηριστικά της εικόνας αντί να απομνημονεύουν συγκεκριμένες κλίμακες απόστασης.

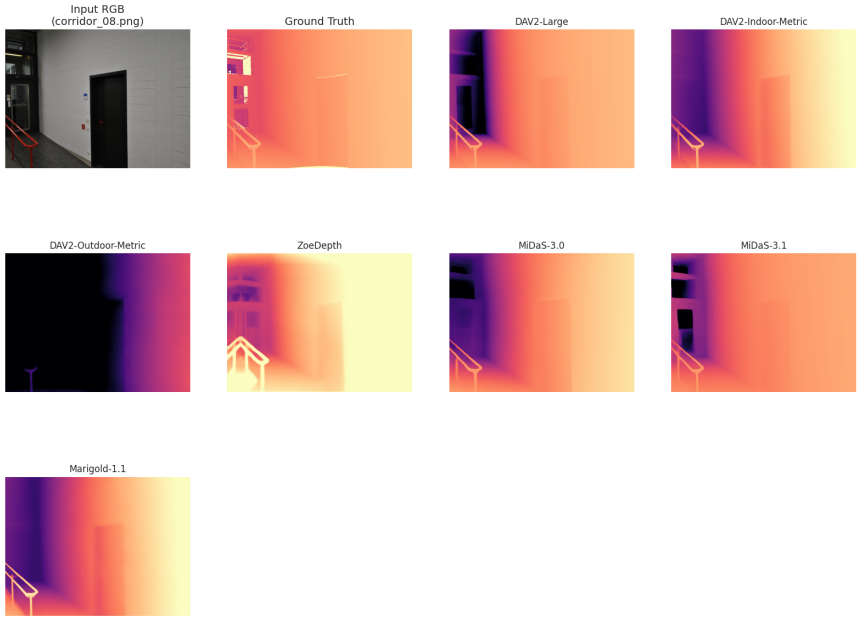
3) Ποιοτική Αξιολόγηση: Η οπτική επιθεώρηση των αποτελεσμάτων, όπως παρουσιάζεται στα Σχήματα 3α′ και 3β′, έρχεται να επιβεβαιώσει και να ερμηνεύσει τις ποσοτικές μετρήσεις, αναδεικνύοντας χαρακτηριστικά που δεν αποτυπώνονται πλήρως στους αριθμητικούς δείκτες.

Στο Σχήμα 3α′, το οποίο απεικονίζει ένα δείγμα από το σύνολο δεδομένων IBims-1, παρατηρείται η ικανότητα του DAV2-Large να διατηρεί ευκρινή τα όρια των αντικειμένων (sharp edges). Σε αντίθεση με τα άλλα μοντέλα που τείνουν να εξομαλύνουν (smooth out) τις λεπτομέρειες, το DAV2 και το Marigold ανακτούν με επιτυχία τη γεωμετρία μικρότερων αντικειμένων, όπως τα πόδια των επίπλων και τα διακοσμητικά στοιχεία. Ειδικότερα, το Marigold, ως μοντέλο διάχυσης, επιδεικνύει εντυπωσιακή λεπτομέρεια στην εντόπιση των ακμών των αντικειμένων. Από την άλλη, είναι εμφανές ότι το παράθυρο στα αριστερά της εικόνας που περιέχει μια αντανάκλαση είναι ένα πολύ δύσκολο σημείο για όλα τα μοντέλα. Αυτός είναι και ένας από τους λόγους επιλογής αυτού του συνόλου δεδομένων.

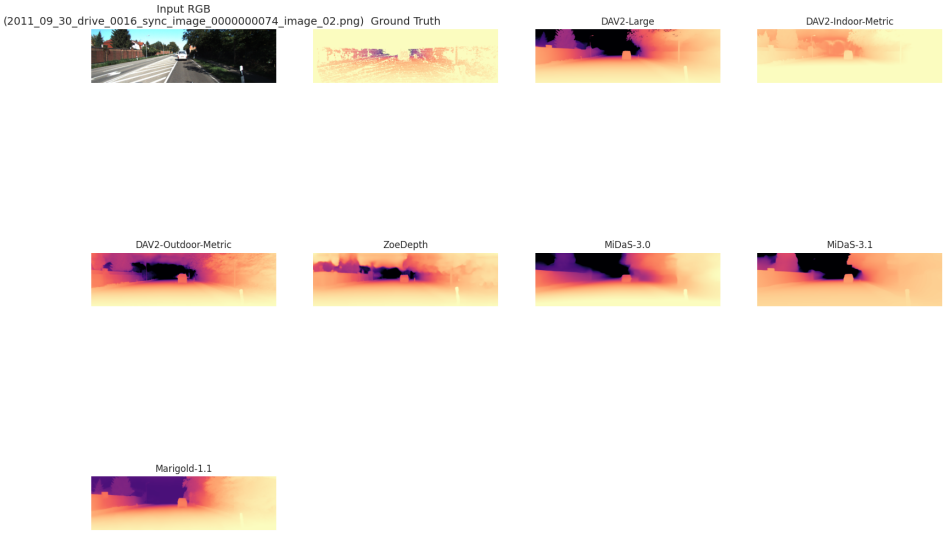
Αντίστοιχα, στο Σχήμα 3β′ (εξωτερικό περιβάλλον από το σύνολο KITTI), η υπεροχή των θεμελιωδών μοντέλων Depth Anything V2 και Marigold καθίσταται σαφής στη διαχείριση του βάθους σε μεγάλες αποστάσεις. Επιτυγχάνει συνεπή διαχωρισμό του δρόμου από τα οχήματα και το φόντο. Παρατηρείται επίσης ότι αυτά τα μοντέλα διαχειρίζονται πολύ πιο αποτελεσματικά τις λεπτές δομές, όπως τους κορμούς των δέντρων και τους στύλους, που αποτελούν παραδοσιακά σημεία αποτυχίας για τους αλγορίθμους εκτίμησης βάθους.

X. Συμπερασματα

Η παρούσα εργασία επιχείρησε μια γενική αξιολόγηση σύγχρονων μοντέλων εκτίμησης βάθους (Monocular Depth Estimation), εστιάζοντας στη δυνατότητα γενίκευσης σε άγνωστα περιβάλλοντα (zero-shot generalization). Μέσα από πειραματική διαδικασία σε τρία ετερογενή σύνολα δεδομένων (IBims-1, NYU Depth V2, KITTI), προέκυψαν σημαντικά συμπεράσματα σχετικά με την εξέλιξη του πεδίου.



(α′) Ποιοτική αξιολόγηση σε εσωτερικό χώρο (IBims-1).



(β′) Ποιοτική αξιολόγηση σε εξωτερικό περιβάλλον (KITTI).

Σχήμα 3: Οπτική σύγκριση των αποτελεσμάτων (Qualitative Evaluation). Τα θεμελιώδη μοντέλα (Depth-Anything-V2, Marigold) παράγουν πιο ευκρινείς και δομικά συνεπείς χάρτες βάθους σε σχέση με παλαιότερες προσεγγίσεις, τόσο σε εσωτερικά όσο και σε εξωτερικά σενάρια.

Πρωτίστως, καταδείχθηκε η σαφής υπεροχή των θεμελιωδών μοντέλων (Foundation Models), όπως το Depth Anything V2, έναντι παλαιότερων ή εξειδικευμένων αρχιτεκτονικών. Η εκπαίδευση σε τεράστια κλίμακα δεδομένων (συμπεριλαμβανομένων συνθετικών εικόνων) επιτρέπει στο μοντέλο να μαθαίνει σημασιολογικά χαρακτηριστικά και όχι απλώς να απομνημονεύει γεωμετρικές σταθερές, επιτυγχάνοντας έτσι εξαιρετική απόδοση τόσο σε εσωτερικούς όσο και σε εξωτερικούς χώρους.

Δεύτερον, η ανάλυση ανέδειξε το trade-off μεταξύ Ακρίβειας και Ταχύτητας. Τα μοντέλα διάχυσης (Diffusion Models), όπως το Marigold, προσφέρουν την υψηλότερη ποιότητα ανακατασκευής λεπτομερειών και υφής, πλησιάζοντας περισσότερο στην ανθρώπινη αντίληψη. Ωστόσο, το υψηλό υπολογιστικό κόστος τους καθιστά προς το παρόν απαγορευτική τη χρήση τους σε εφαρμογές πραγματικού χρόνου (real-time), σε αντίθεση με τα πιο αποδοτικά μοντέλα που περιέχουν transformers (πχ DAV2, ZoeDepth).

Τρίτον, επιβεβαιώθηκε ότι το χάσμα πεδίου (Domain Gap) παραμένει το μεγάλο πρόβλημα των μοντέλων που στοχεύουν απευθείας σε μετρικό βάθος (Metric Depth) χωρίς εκτεταμένη προ-εκπαίδευση. Η προσέγγιση του σχετικού βάθους (Relative Depth) με μεταγενέστερη ευθυγράμμιση αποδεικνύεται πιο ασφαλής στρατηγική για εφαρμογές γενικού σκοπού ("in-the-wild").

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

[11] T. Koch, L. Liebel, F. Fraundorfer, and M. Körner, ``Evaluation of cnn-based single-image depth estimation methods,'' in *Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS)*, S. Leal-Taixé, Laura;Roth, Ed. Springer International Publishing, 2019, pp. 331--348. [Online]. Available: http://openaccess.thecvf.com/content_ECCVW_2018/papers/11131/Koch_Evaluation_of_CNN-based_Single-Image_Depth_Estimation_Methods_ECCVW_2018_paper.pdf

[12] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, ``Indoor segmentation and support inference from rgbd images,'' in *ECCV*, 2012.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, ``Vision meets robotics: The kitti dataset,'' *International Journal of Robotics Research (IJRR)*, 2013.