**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**

Σχολή Αγρονόμων και Τοπογράφων Μηχανικών - Μηχανικών Γεωπληροφορικής

Εργαστήριο Τηλεπισκόπησης

---

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

# ΑΦΑΙΡΕΣΗ ΣΥΝΝΕΦΩΝ ΑΠΟ ΧΡΟΝΟΣΕΙΡΕΣ ΔΟΡΥΦΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ SENTINEL ΜΕ ΤΗ ΧΡΗΣΗ ΔΙΚΤΥΩΝ AXIAL TRANSFORMERS

---

ΧΡΙΣΤΟΠΟΥΛΟΣ ΔΙΟΝΥΣΗΣ

ΙΟΥΝΙΟΣ 2022

ΑΘΗΝΑ

**NATIONAL TECHNICAL UNIVERSITY OF ATHENS**

School of Rural, Surveying and Geoinformatics Engineering

Remote Sensing Lab

DIPLOMA THESIS

# CLOUD REMOVAL FROM MULTITEMPORAL SATELLITE IMAGES USING AXIAL TRANSFORMER NETWORKS

CHRISTOPOULOS DIONYSIS

JUNE 2022

ATHENS

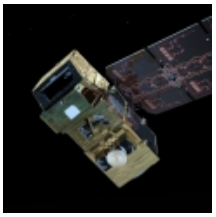
RSLab

Remote Sensing Laboratory

National Technical University of Athens

✓ Sensing ✓ Analytics ✓ Monitoring

# Περίληψη

Η αφαίρεση των συννέφων από δορυφορικά πολυφασματικά δεδομένα είναι ένα κρίσιμο ζήτημα που απασχολεί τον τομέα της Τηλεπισκόπησης, ειδικά με σκοπό την αναδόμηση εικόνων που καλύπτουν περιοχές ενδιαφέροντος. Στην εποχή μας, τέτοιου είδους δεδομένα υψηλής χωρικής και χρονικής ανάλυσης είναι ευρέως διαθέσιμα με τη βοήθεια πολυάριθμων δορυφόρων οι οποίοι βρίσκονται σε τροχιά στο διάστημα και παρακολουθούν συνεχώς τη Γη. Το πρόβλημα έγκειται στο γεγονός ότι μεγάλο πλήθος των συλλεγόμενων δεδομένων δεν μπορεί να χρησιμοποιηθεί καθώς επηρεάζονται απο σύννεφα. Ο βαθμός επιρροής των συννέφων διαφέρει ανά περιοχή και εξαρτάται από τη γεωγραφική της θέση, το κλίμα της και την εποχή στην οποία λαμβάνονται οι παρατηρήσεις. Οι μέθοδοι αφαίρεσης συννέφων που έχουν αναπτυχθεί χωρίζονται σε δύο κατηγορίες, βάση της προσέγγισης του προβλήματος, δηλαδή σε εκείνες που εκμεταλλεύονται τη διαχρονική εξέλιξη των τιμών κάθε εικονοστοστοιχείου και εκείνων που επιχειρούν να γεμίσουν τα κενά βασιζόμενες αποκλειστικά σε μια εικόνα. Οι πρώτες πλεονεκτούν καθώς λαμβάνουν υπόψιν την χρονική μεταβολή των τιμών κάθε εικονοστοιχείου μειώνοντας την αβεβαιότητα πρόβλεψης των τιμών που λείπουν. Διάφορες μέθοδοι βαθιάς μάθησης έχουν αναπτυχθεί για την αφαίρεση νέφους από μια εικόνα (π.χ [7]), ωστόσο η καταγραφή τόσο των χωρικών όσο και των χρονικών σχέσεων οδηγεί σε βελτιωμένα αποτελέσματα. Προτείνουμε, λοιπόν, μια καινοτόμα μέθοδο αφαίρεσης συννέφων από χρονοσειρές δορυφορικών εικόνων η οποία βασίζεται στα δίκτυα των «Transformers» [6] που χρησιμοποιούν το μηχανισμό του attention (προσοχής) τόσο στη χωρική όσο και στη χρονική διάσταση.



**Λέξεις-Κλειδιά:** χρονοσειρές, sentinel-2, αναδόμηση, αφαίρεση συννέφων, βαθιά μάθηση, axial attention



RSLab
Remote Sensing Laboratory
National Technical University of Athens
✓ Sensing ✓ Analytics ✓ Monitoring

# Abstract

Nowadays, EO data[1] of high spatial and temporal resolution are available thanks to the commissioning of a large fleet of satellites continuously monitoring the earth. However, a large proportion of the collected data cannot be used as they are affected by clouds. This proportion depends on the geographic location of the region, its climate characteristics, and the season. Cloud removal methods can be divided in two main categories, namely, those taking advantage of the temporal evolution of the pixel values and those attempting inpainting/gap-filling of parts affected by clouds in a single image. The first ones have the advantage of being conditioned on the history of each pixel, reducing the prediction uncertainty of the missing values. A few deep learning methods have been developed for single image cloud removal, based on generative neural networks (e.g. [7]). Capturing both the temporal and spatial relations of the pixels, although more challenging, can lead to improved results. We propose a novel method for cloud removal from multitemporal satellite images based on transformer networks [6] that use the attention mechanism both in the spatial and temporal dimensions.

**Keywords:** time-series, sentinel-2, reconstruction, cloud-free, deep learning, axial attention, autoregressive models

RSLab
Remote Sensing Laboratory
National Technical University of Athens
√ Sensing √ Analytics √ Monitoring

[1] Earth Observation data

# Ευχαριστίες

Με την εκπόνηση αυτής της διπλωματικής εργασίας θέλω να ευχαριστήσω θερμά τον επιβλέπων καθηγητή Κωνσταντίνο Καράντζαλο για τις συμβουλές, τη στήριξη και την εμπιστοσύνη που μου έδειξε όλον αυτό τον καιρό. Στη συνέχεια, θέλω να ευχαριστήσω τον κ. Μάκη Ντούσκο για την καθοριστική βοήθειά του σε όλη την πορεία της εργασίας και τις πολλές ώρες που αφιερώσαμε για την επίτευξη αυτού του εξαιρετικού αποτελέσματος. Μέσω του πολύ ωραίου κλίματος και της άψογης συνεργασίας επιτεύχθηκε η δημοσίευση της παρούσας εφαρμογής στο συνέδριο ISPRS 2022 (International Society for Photogrammetry and Remote Sensing).

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένεια μου, που δεν σταμάτησαν ποτέ να πιστεύουν σε εμένα και για την στήριξη τους όλα αυτά τα χρόνια σπουδών.

RSLab
Remote Sensing Laboratory
National Technical University of Athens
✓ Sensing ✓ Analytics ✓ Monitoring

# Περιεχόμενα

# ΚΕΦΑΛΑΙΟ 1

# Εισαγωγή

## 1.1 Περιγραφή Στόχου

Διαχρονικό πρόβλημα στους τομείς της φωτοερμηνείας και τηλεπισκόπησης αποτελεί η μαζική ύπαρξη συννέφων στα δορυφορικά πολυφασματικά δεδομένα, η οποία αποτρέπει τους ενδιαφερόμενους από την μελέτη περιοχών ενδιαφέροντος. Πολυάριθμες και διαφορετικές προσεγγίσεις έχουν αναπτυχθεί ανά τα έτη για την αντιμετώπιση του συγκεκριμένου προβλήματος, είτε με τη χρήση βαθιάς μηχανικής μάθησης ή χωρίς αυτή (*median*).

Στη παρούσα εργασία θα μελετήσουμε κυρίως τις μεθόδους βαθιάς μηχανικής μάθησης καθώς τα αποτελέσματα τους προσφέρουν σημαντικά αυξημένες ακριβείες συγκριτικά με απλούστερες προσεγγίσεις. Με τη σειρά τους ωστόσο, οι τεχνικές βαθιάς μάθησης που έχουν προταθεί με σκοπό την αφαίρεση των νεφών από δορυφορικές εικόνες διαχωρίζονται σε δύο επιμέρους κατηγορίες. Την πρώτη κατηγορία απαρτίζουν οι προσεγγίσεις οι οποίες βασίζονται αποκλειστικά σε μια εικόνα, επεξεργάζονται δηλαδή την πληροφορία που προσφέρει μια μοναδική χρονική στιγμή[1] και επιχειρούν ένα χρωματικό γέμισμα των κενών που προέρχονται από τα σύννεφα.

Στις μέρες μας, με τον συνεχώς αυξανόμενο στόλο δορυφόρων που βρίσκονται σε τροχιά γύρω από τη Γη, υπάρχει η δυνατότητα λήψης εικόνων της ίδιας περιοχής σε πολλές χρονικές στιγμές με μεγάλη συχνότητα, η οποία εξαρτάται από τον εκάστοτε δορυφόρο. Όπως είναι σαφές, λοιπόν, στη δεύτερη κατηγορία ανήκουν οι προσεγγίσεις που βασίζονται στη διαχρονική εξέλιξη κάθε εικονοστοιχείου, οπότε αντιμετωπίζουν το πρόβλημα λαμβάνοντας υπόψιν χωρικές και χρονικές σχέσεις[2]. Αναλυτικότερη μελέτη και σύγκριση των μεθόδων πραγματοποιείται στο Κεφάλαιο 3.

Η μέθοδος που προτείνεται σε αυτή την εργασία, χρησιμοποιεί τον άξονα του χρόνου ώστε τα εικονοστοιχεία, επηρεασμένα από σύννεφα, που τίθενται προς εξέταση, να αναδομούνται

[1] Ενδεικτικές προσεγγίσεις: CloudGAN[7], SpAGAN[13]
[2] Ενδεικτικές προσεγγίσεις: FFVI[9], Fill&Fit[16]

βάση των τιμών και της δομής που παρουσιάζουν σε προηγούμενες χρονικές στιγμές. Για να επιτευχθεί το παραπάνω ζητούμενο επικαλούμαστε την αρχιτεκτονική των **Transformers**[6] σε συνδυασμό με τη τεχνική του **Axial Attention**[10] για μειωμένο υπολογιστικό κόστος.

## 1.2 Βασικές Έννοιες

Σε αυτή την ενότητα θα καλύψουμε μερικές από τις σημαντικότερες έννοιες, απλές και σύνθετες, που θα συναντήσουμε στη διάρκεια της εργασίας και θα βοηθήσουν τον αναγνώστη στη πλήρη κατανόηση της.

**Τηλεπισκόπηση (Remote Sensing):** η επιστήμη συλλογής δεδομένων για κάποιο αντικείμενο απο απόσταση. Η Αμερικανική Εταιρία Φωτογραμμετρίας και Τηλεπισκόπησης (*ASPRS - American Society for Photogrammetry and Remote Sensing*) δίνει τον επίσημο ορισμό ως εξής:

> *«Η μέτρηση ή συλλογή πληροφοριών για κάποια ιδιότητα ενός αντικειμένου ή φαινομένου μέσω κάποιου οργάνου καταγραφής, το οποίο δεν βρίσκεται σε άμεση επαφή με το υπό μελέτη αντικείμενο ή φαινόμενο»*
> (Colwell, 1983)

**Αισθητήρας (sensor):** το όργανο με τη βοήθεια του οποίου πραγματοποιείται η *Τηλεπισκόπηση*. Οι περισσότεροι αισθητήρες καταγράφουν την Ηλεκτρομαγνητική Ακτινοβολία (*EMR - Electromagnetic Radiation*) η οποία μεταδίδεται από την πηγή με ταχύτητα $3 \times 10^8\ m \cdot s^{-1}$, είτε από το κενό του διαστήματος είτε από ανάκλαση ή επανεκπομπή προς τον αισθητήρα. Οι μεταβολές στην ποσότητα και τις ιδιότητες της ηλεκτρομαγνητικής ακτινοβολίας, όταν εντοπίζονται από τον αισθητήρα, αποτελούν πολύτιμη πηγή δεδομένων για την ερμηνεία σημαντικών ιδιοτήτων του φαινομένου (π.χ. θερμοκρασία, χρώμα).

**Πλεονεκτήματα και περιορισμοί:** Η τηλεπισκόπηση δεν είναι παρεμβατική επιστήμη, καθώς η απλή καταγραφή της ηλεκτρομαγνητικής ενέργειας, που ανακλάται ή εκπέμπεται, δεν επηρεάζει το υπό εξέταση φαινόμενο ή αντικείμενο. Προσφέρει συστηματική, μαζική και ταχύτατη συλλογή δεδομένων με εξαιρετικά χαμηλό κόστος, εξαλείφοντας τυχόν δειγματοληπτικά σφάλματα που προκύπτουν από επιτόπιες έρευνες. Τέλος, τα τηλεπισκοπικά δεδομένα παρέχουν σημαντικές επιστημονικές πληροφορίες, σε κάθε λήψη, για το περιβάλλον, το κλίμα και τα χαρακτηριστικά της υπο εξέτασης περιοχής ή φαινομένου όπως για παράδειγμα *συντεταγμένες (x, y), υψόμετρο, βάθος, βιομάζα, θερμοκρασία, περιεκτικότητα σε υγρασία*. Ο μεγαλύτερος περιορισμός της Τηλεπισκόπησης, είναι το γεγονός πως σε αρκετές περιπτώσεις δεν προσφέρει τις απαραίτητες πληροφορίες με αποτέλεσμα να απαιτείται και επιτόπια έρευνα για την πλήρη λήψη των δεδομένων.

**Φασματική Διακριτική Ικανότητα (spectral resolution):** αναφέρεται στον αριθμό και τη διάσταση συγκεκριμένων διαστημάτων μήκους κύματος στο ηλεκτρομαγνητικό φάσμα, (κανάλια ή ζώνες) στα οποία είναι ευαίσθητο ένα όργανο τηλεπισκόπησης. Τα πολυφασματικά (multispectral) συστήματα τηλεπισκόπησης καταγράφουν ενέργεια σε πολλά κανάλια του ηλεκτρομαγνητικού φάσματος.

**Χωρική Διακριτική Ικανότητα (spatial resolution):** είναι το μέτρο της μικρότερης γωνιακής ή γραμμικής απόστασης ανάμεσα σε δύο αντικείμενα που μπορεί να διακριθεί από το σύστημα τηλεπισκόπησης.

**Χρονική Διακριτική Ικανότητα:** αναφέρεται στο πόσο συχνά ο αισθητήρας καταγράφει εικόνες από μια συγκεκριμένη περιοχή.

**Ραδιομετρική Διακριτική Ικανότητα (radiometric resolution):** η ευαισθησία ενός τηλεπισκοπικού ανιχνευτή σε διαφορές στην ένταση των σημάτων που λαμβάνει, καθώς καταγράφει την ισχύ της ακτινοβολίας που εκπέμπεται, ανακλάται ή οπισθοσκεδάζεται από το έδαφος. Καθορίζει, λοιπόν, τον αριθμό των διακριτών επιπέδων σήματος όπως φαίνεται και στον Πίνακα 1.1.

| Bit-depth | Εύρος |
|---|---|
| 7-bits | (0-127) |
| 8-bits | (0-255) |
| 9-bits | (0-511) |
| 10-bits | (0-1023) |

**Πίνακας 1.1:** Ραδιομετρική Διακριτική Ικανότητα.

**Μηχανική Μάθηση (Machine Learning):** η επιστήμη που διερευνά τον σχεδιασμό και την κατασκευή μοντέλων τα οποία μέσω της χρήσης κατάλληλων αλγορίθμων, προσεγγίζουν (μαθαίνουν) μια (άγνωστη) συνάρτηση/κατανομή ενδιαφέροντος, απευθείας από δεδομένα που τους παρέχονται και να λαμβάνουν αποφάσεις ανάλογα με τον στόχο που τους ανατίθεται. Κατά την εφαρμογή τους, ο ανθρώπινος παράγοντας περιορίζεται αποκλειστικά στην τροφοδοσία των δεδομένων. Στόχος της μηχανικής μάθησης, λοιπόν, είναι η κατανόηση της δομής των δεδομένων και η προσαρμογή θεωρητικών συναρτήσεων/κατανομών πάνω σε αυτά. Για το σκοπό αυτό, συνήθως, χρησιμοποιούνται επαναληπτικές προσεγγίσεις έως ότου ο αλγόριθμος «μάθει» ένα ισχυρό μοτίβο που να προσαρμόζεται στα δεδομένα.

Η μηχανική μάθηση χωρίζεται σε τέσσερις μεγάλες κατηγορίες ανάλογα την φύση, τη δομή των

δεδομένων εκπαίδευσης ή την ανατροφοδότηση/επίβλεψη του μοντέλου:

- Επιβλεπόμενη Μάθηση (Supervised): Η μηχανή δέχεται δεδομένα εκπαίδευσης με τα αντίστοιχα επιθυμητά αποτελέσματα. Στόχος της είναι να μάθει την συνάρτηση/κατανομή (κανόνα) που συνδέει τα δεδομένα εισόδου με τα επιθυμητά αποτελέσματα. Χαρακτηριστικές εφαρμογές της επιβλεπόμενης μάθησης είναι η ταξινόμηση, η παλινδρόμηση και η πρόβλεψη μελλοντικών καταστάσεων.
- Μη-Επιβλεπόμενη Μάθηση (Unsupervised): Η μηχανή δέχεται δεδομένα εκπαίδευσης δίχως τα επιθυμητά αποτελέσματα. Στόχος της είναι να μάθει από μόνη της τη δομή και τις σχέσεις που χαρακτηρίζουν τα δεδομένα εισόδου.
- Ημι-Επιβλεπόμενη Μάθηση (Semi-supervised): Αποτελεί συνδυασμό των δύο προηγούμενων προσεγγίσεων. Συνήθως η ημι-επιβλεπόμενη μάθηση δέχεται μαζί με τις εισόδους και έναν μικρό μέρος των επιθμητών αποτελεσμάτων. Η μέθοδος αυτή χρησιμοποιείται όταν το κόστος για την τροφοδότηση του αλγορίθμου εκπαίδευσης, με όλα τα επιθυμητά αποτελέσματα, είναι πολύ μεγάλο. Μπορεί να χρησιμοποιηθεί για τις ίδιες εφαρμογές με την επιβλεπόμενη μάθηση.
- Ενισχυτική Μάθηση (Reinforced): Ο αλγόριθμος αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτευχθεί ένας συγκεκριμένος στόχος. Μέσω δοκιμών και λάθους και με ένα σύστημα ανταμοιβών μαθαίνει τη σωστή διαδικασία για την επίτευξη αυτού του στόχου. Οι βασικότερες χρήσεις τις ενισχυτικής μάθησης βρίσκονται στη ρομποτική, στον τομέα των παιχνιδιών και στην πλοήγηση.

**Βαθιά Μάθηση (Deep Learning):** Αποτελεί υποκατηγορία της Μηχανικής Μάθησης, διαθέτει πολυπλοκότερα μοντέλα, με μεγαλύτερο «βάθος» και χρησιμοποιεί πολλά επίπεδα με σκοπό τη σταδιακή εξαγωγή χαρακτηριστικών υψηλότερου επιπέδου.

Αξίζει σε αυτό το σημείο να αναφερθεί, πως η προσέγγιση που παρουσιάζεται σε αυτή την εργασία ανήκει στην κατηγορία της Βαθιάς μάθησης, καθώς χρησιμοποιεί το δίκτυο του Axial Transformer που με την σειρά του είναι βασισμένο στην τεχνική του «Axial Attention» η οποία αναλύεται εκτενώς στο Κεφάλαιο 2.

**Συνελικτικά επίπεδα (Convolutional Layers):** Χρησιμοποιούνται συνήθως σε προβλήματα που αφορούν την επεξεργασία εικόνων. Στο επίπεδο αυτό γίνονται πράξεις, γνωστές ως συνελίξεις, μεταξύ της εικόνας εισόδου και μιας σειράς φίλτρων (τετραγωνικών, συνήθως, πινάκων μικρών διαστάσεων), με σκοπό την εξαγωγή χαρακτηριστικών μοτίβων της εικόνας. Κάθε φίλτρο είναι ένας δισδιάστατος πίνακας, μεγέθους $3\times3$ ή $5\times5$ στην πλειοψηφία των περιπτώσεων, ο οποίος εφαρμόζεται σε ένα αντίστοιχο μέρος της εικόνας και υπολογίζεται το εσωτερικό γινόμενο μεταξύ των εικονοστοιχείων και των βαρών που απαρτίζουν το φίλτρο. Η διαδικασία αυτή

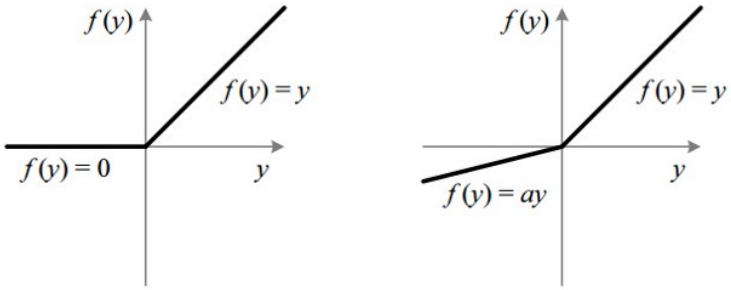
επαναλαμβάνεται με μετάθεση του φίλτρου, βάση καθορισμένου βήματος, ώσπου να καλύψει όλα τα εικονοστοιχεία. Οι τιμές των βαρών των φίλτρων εκτιμώνται κατά την διαδικασία της εκπαίδευσης του δικτύου.

Μερικές σημαντικές παράμετροι των συνελικτικών στρώσεων που πρέπει να ληφθούν υπόψιν είναι:

- Το μέγεθος του φίλτρου. (1×1, 3×3, 5×5, κ.α)
- Ο αριθμός των φίλτρων κάθε στρώσης, ο οποίος επηρεάζει άμεσα το βάθος της εξόδου.
- Το βήμα (stride) μετάθεσης του φίλτρου (σε εικονοστοιχεία) ύστερα από κάθε συνέλιξη. Ορίζεται συνήθως 1, ωστόσο για μεγαλύτερες τιμές τόσο μικρότερες είναι οι διαστάσεις της εξόδου.
- Η επέκταση (padding), καλείται όταν το μέγεθος του φίλτρου σε συνδυασμό με της εικόνας δεν του επιτρέπει να εφαρμοστεί σε όλα τα εικονοστοιχεία. Τα μέρη της εικόνας που κατά την εφαρμογή του φίλτρου βρίσκονται έξω από τα όρια της θεωρούνται (συνήθως) ότι έχουν μηδενικές τιμές.

**Μη-γραμμικές Συναρτήσεις Ενεργοποίησης (Activation Functions):** Κάθε συνελικτικό επίπεδο ακολουθείται από μια συνάρτηση ενεργοποίησης. Χωρίζονται σε τρεις μεγάλες κατηγορίες, τις δυαδικές, τις γραμμικές και τις μη-γραμμικές. Οι μη γραμμικές συναρτήσεις ενεργοποίησης συμπεριλαμβάνονται στα νευρωνικά δίκτυα με σκοπό την εκμάθηση πολύπλοκων μοτίβων για τα δεδομένα. Στην ουσία, αποφασίζουν εάν ένας νευρώνας θα ενεργοποιηθεί και αν περιέχει σημαντική πληροφορία, η οποία πρέπει να τροφοδοτηθεί στα επόμενα επίπεδα. Μερικές σημαντικές μη-γραμμικές συναρτήσεις ενεργοποίησης που αναφέρονται σε αυτή την εργασία είναι οι εξής:

- **ReLU (Rectified Linear Unit):** Η συνάρτηση αυτή επιστρέφει την τιμή 0 αν η είσοδος της είναι αρνητική, ενώ για οποιαδήποτε θετική τιμή $x$ επιστρέφει το ίδιο το $x$, ακολουθώντας την εξίσωση $f(x) = max(0, x)$. Το εύρος των τιμών της είναι $[0, +\infty]$. Ο βασικός περιορισμός της ReLU είναι πως για κάθε αρνητική τιμή εισόδου η κλίση μηδενίζεται, γεγονός το οποίο αποτρέπει τα βάρη να ανανεωθούν και να προσαρμοστούν κατάλληλα, οπότε και οι αντίστοιχοι νευρώνες αδρανοποιούνται πλήρως.
- **Leaky ReLU:** Αποτελεί παραλλαγή της συνάρτησης ReLU, η οποία διορθώνει ως έναν βαθμό το πρόβλημα των αδρανών νευρώνων που προαναφέρθηκε, εφαρμόζοντας μια μικρή σταθερή θετική κλίση. Ακολουθεί την εξίσωση $f(x) = max(ax, x)$ όπου συνήθως $a = 0.01$. Το εύρος των τιμών της είναι $[-\infty, +\infty]$ και με αυτόν τον τρόπο, κάθε αρνητική είσοδος αποκτά μη μηδενική κλίση, οπότε οι αντίστοιχοι νευρώνες δεν αδρανοποιούνται.




**Σχήμα 1.1:** Αριστερά: Συνάρτηση ενεργοποίησης ReLU, Δεξιά: Συνάρτηση ενεργοποίησης Leaky ReLU.

- **Softmax:** Χρησιμοποιείται συνήθως σε προβλήματα ταξινόμησης πολλαπλών κατηγοριών. Έξοδος της συνάρτησης είναι μια κατανομή πιθανοτήτων $\sigma \in [0, 1]$ για κάθε πιθανή κατηγορία του μοντέλου.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{1.1}$$

  όπου $z_i$ το διάνυσμα εισόδου, $z_j$ το διάνυσμα εξόδου και $N$ το πλήθος των πιθανών κατηγοριών.

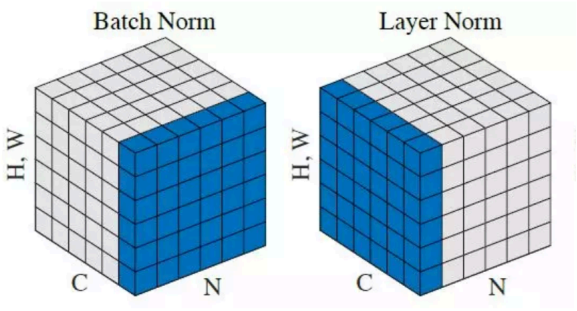
**Batch Normalization:** Χρησιμοποιείται αμέσως πριν ή αμέσως μετά τις συναρτήσεις ενεργοποίησης με σκοπό την κανονικοποίηση των εισόδων που προέρχονται από προηγούμενα επίπεδα του δικτύου. Προσαρμόζει τις εισόδους ώστε να έχουν μηδενικό μέσο όρο και τυπική απόκλιση ίση με τη μονάδα και οδηγεί το δίκτυο σε ταχύτερους χρόνους εκπαίδευσης, όντας παράλληλα και πιο σταθερό. Τέλος, εφαρμόζει δύο μετασχηματισμούς κλίμακας και μετάθεσης, στα κανονικοποιημένα δεδομένα, μέσω των εκπαιδεύσιμων παραμέτρων $\gamma$ και $\beta$ αντίστοιχα. Οι παραπάνω διεργασίες που περιγράφονται στις εξισώσεις 1.2 - 1.5 πραγματοποιούνται σε πολλές ταυτόχρονες υποπεριοχές των εισόδων του δικτύου (mini-batches) και όχι σε κάθε είσοδο ξεχωριστά.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.2}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{1.3}$$

$$x_{i,norm} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{1.4}$$

$$BN(x_i) = \gamma \cdot x_{i,norm} + \beta \tag{1.5}$$




**Σχήμα 1.2:** Batch Normalization - Layer Normalization

όπου $n$ το μέγεθος του mini-batch, $\mu, \sigma$ ο μέσος όρος και η τυπική απόκλιση των εξεταζόμενων στοιχείων του mini-batch, $\epsilon$ ένας σταθερός αριθμητικός παράγοντας για την αποφυγή διαίρεσης με μηδενικό παρανομαστή και $\gamma, \beta$ οι εκπαιδεύσιμες παράμετροι του επιπέδου.

**Layer Normalization:** Είναι και αυτό με τη σειρά του ένα επίπεδο κανονικοποίησης των εισόδων, αλλά σε αυτή την περίπτωση οι υπολογισμοί γίνονται στον άξονα των χαρακτηριστικών κάθε μιας ξεχωριστά. Δεν δημιουργούνται λοιπόν νέες αλληλεξαρτήσεις μεταξύ των διαφορετικών εισόδων όπως συμβαίνει στο επίπεδο του Batch Normalization. Η τεχνική αυτή, χρησιμοποιείται και στους Transformers που θα μελετήσουμε στο Κεφάλαιο 2.

**Αλγόριμος Βελτιστοποίησης RMSProp:** Χρησιμοποιείται στην εκπαίδευση πολύπλοκων νευρωνικών δικτύων. Στον υπολογισμό των κλίσεων, μέσω των παραγώγων, κατά την «προς τα πίσω μετάδοση» της πληροφορίας στο δίκτυο, παρουσιάζεται συχνά το πρόβλημα εξαφάνισης τους. Ο αλγόριθμος βελτιστοποίησης RMSProp αντιμετωπίζει το παραπάνω ζήτημα με την προσαρμογή του συντελεστή εκμάθησης (learning rate). Η αλλαγή των βαρών οδηγεί σε διαφορετικά μεγέθη κλίσεων, δυσκολεύοντας την εύρεση ενός ενιαίου συντελεστή εκπαίδευσης για όλο το δίκτυο, οπότε ο RMSProp με ένα κινούμενο παράθυρο μέσου όρου του τετραγώνου των κλίσεων προσαρμόζει τις μεταβολες αυτές.

$$E[g^2]_t = \beta E[g^2]_t + (1-\beta) g_t^2 \tag{1.6}$$

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{1.7}$$

όπου $E[g^2]_t$ ο κινούμενος μέσος όρος του τετραγώνου των κλίσεων, $g_t$ οι κλίσεις της συνάρτησης κόστους σε σχέση με το βάρος $w$, $\eta$ η αρχικοποίηση του συντελεστή εκμάθησης και $\beta$ αριθμητική

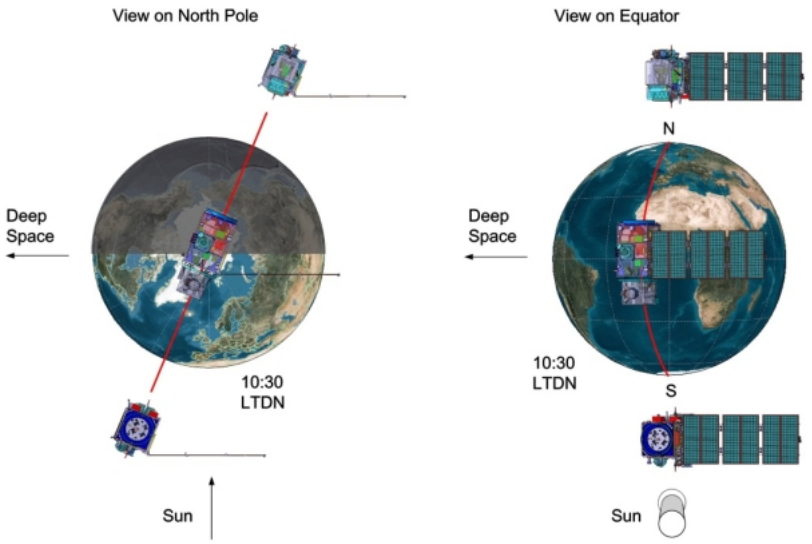
παράμετρους του κινούμενου μέσου όρου με προτεινόμενη τιμή $0.9$.

## 1.3 Πολυφασματικά Δεδομένα Sentinel-2

Στην ενότητα αυτή θα ασχοληθούμε με την αποστολή Sentinel-2 από την οποία έχουν αντληθεί τα δεδομένα για τα πειράματα της εργασίας.

### 1.3.1 Γενικά Στοιχεία Αποστολής

Η αποστολή Sentinel-2 είναι μέρος του προγράμματος Copernicus και παρέχει, υψηλής ανάλυσης, πολυφασματικά δεδομένα. Αποτελείται από έναν σχηματισμό δύο «δίδυμων» ηλιοσύγχρονων δορυφόρων, τους Sentinel-2A & Sentinel-2B, οι οποίοι βρίσκονται στην ίδια τροχιά στο διάστημα, αλλά με διαφορά φάσης $180^{\circ}$ μεταξύ τους. Με τον σχηματισμό αυτό (Σχήμα 1.3), η αποστολή έχει συχνότητα επανεπισκεψιμότητας 5 ημέρες.




**Σχήμα 1.3:** Η τροχιά των δίδυμων δορυφόρων της αποστολής Sentinel-2.

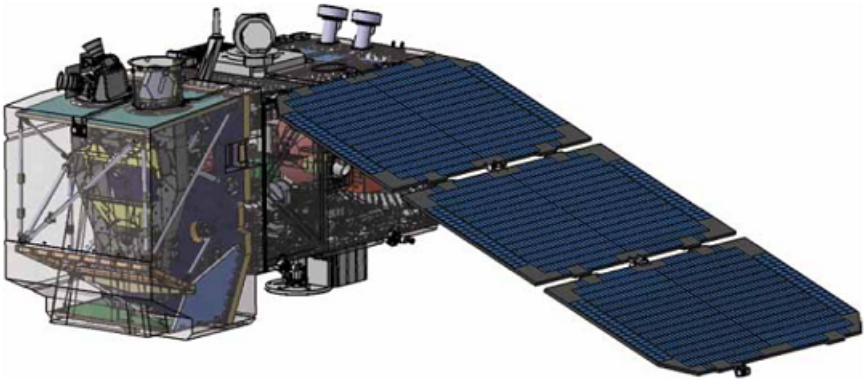
Οι δύο δορυφόροι έχουν ενσωματωμένο αισθητήρα με ικανότητα ανίχνευσης και διαχωρισμού 13 φασματικών ζωνών: τέσσερις με χωρική διακριτική ικανότητα **10m**, έξι με **20m** και τρεις με **60m**. Τέλος το πλάτος της τροχιακής λωρίδας, δηλαδή το οπτικό πεδίο είναι στα 290m.

Τα δεδομένα που παρέχονται από την αποστολή προορίζονται για πλήθος εφαρμογών όπως η παρακολούθηση του εδάφους και των μεταβολών στις χρήσεις Γης, η ανάλυση και η διαχείριση κινδύνων (πλημμύρες, σεισμοί), θέματα ασφαλείας (επιτήρηση συνόρων και θαλάσσιων εκτάσεων), η μελέτη της κλιματικής αλλαγής και τέλος η ανάλυση και χαρτογράφηση μεγάλων ωκεάνιων περιοχών[3].

[3] Αναλυτικά στοιχεία για κάθε αποστολή στον επίσημο ιστότοπο της Ευρωπαϊκής Υπηρεσίας Διαστήματος

### 1.3.2 Περιγραφή Δορυφόρων

Κάθε δορυφόρος ζυγίζει περίπου 1,2 τόνους και έχει εκτοξευθεί μέσω του Ευρωπαϊκού εκτοξευτή VEGA. Ο Sentinel-2A εκτοξεύθηκε τον Ιούνιο του 2015 ενώ δύο χρόνια αργότερα, τον Μάρτιο του 2017, ακολούθησε και ο Sentinel-2B. Η προβλεπόμενη διάρκεια ζωής τους είναι 7,25 χρόνια, ωστόσο οι μπαταρίες και τα προωθητικά καύσιμα με τα οποία έχουν εξοπλιστεί, προορίζονται για 12 έτη λειτουργίας, συμπεριλαμβανομένων των ελιγμών εκτός τροχιάς στο τέλος της ζωής τους.



**Σχήμα 1.4:** Sentinel-2 δορυφόρος.

Όπως προαναφέρθηκε, οι δύο δορυφόροι βρίσκονται στην ίδια, ηλιο-σύγχρονη τροχιά με διαφορά φάσης $180^\circ$ και μέσο υψόμετρο 786km. Η τροχιά είναι σύγχρονη με τον ήλιο, ώστε να διασφαλιστεί ότι η γωνία του ηλιακού φωτός στην επιφάνεια της Γης παραμένει σταθερή. Ελαχιστοποιείται κατά το δυνατόν η επίδραση των σκιών και των διαφόρων εντάσεων ακτινοβολίας στο έδαφος, με εξαίρεση τις εποχικές διακυμάνσεις που αποτελούν ένα αναπόφευκτο φαινόμενο. Η μέθοδος αυτή, λοιπόν, διασφαλίζει τη χρονική συνέπεια των δεδομένων και μας επιτρέπει να συλλέγουμε **χρονοσειρές**, οι οποίες αποτελούν και τον θεμέλιο λίθο της παρούσας εργασίας.

### 1.3.3 Πολυφασματικό Όργανο - MSI

Ο αισθητήρας χρησιμοποιεί την τεχνική *push-broom*, συλλέγοντας σειρές από πολυφασματικά δεδομένα κατά μήκος της τροχιακής λωρίδας. Χρησιμοποιεί την κίνηση του δορυφόρου πάνω στη τροχιά ώστε να λαμβάνει συνεχώς νέες σειρές δεδομένων, με μέση περίοδο παρατήρησης τα 17 λεπτά και μέγιστη περίοδο 32 λεπτά, τόσο για χερσαίες όσο και για παράκτιες περιοχές.

Το φως ανακλάται από τη Γη και την ατμόσφαιρά της, προς τον αισθητήρα MSI, ο οποίος λειτουργεί παθητικά. Μέσω ενός φίλτρου, η εισερχόμενη δέσμη φωτός διαχωρίζεται σε δύο συγκροτήματα εστιακού επιπέδου, ένα για τις ορατές και τις κοντινές υπέρυθρες ζώνες (VNIR[4]) και ένα για τις ζώνες υπέρυθρων βραχέων κυμάτων (SWIR[5]). Αμφότερα τα δύο συγκροτήματα, διαθέτουν δύο συστοιχίες 12 ανιχνευτών, τοποθετημένων κλιμακωτά, ώστε να καλύπτουν ολόκληρο το οπτικό πεδίο των 290km στο έδαφος. Περαιτέρω διαχωρισμός στα 13 επιμέρους κανάλια επιτυγχάνεται μέσω φίλτρων που επικαλύπτουν τους ανιχνευτές.

| Χωρική Διακριτική Ικανότητα (m) | Κανάλι | S2A | | S2B | |
|---|---|---|---|---|---|
| | | Μέσο Μήκος Κύματος (nm) | Εύρος (nm) | Μέσο Μήκος Κύματος (nm) | Εύρος (nm) |
| **10** | 2 | 492.4 | 66 | 492.1 | 66 |
| | 3 | 559.8 | 36 | 559.0 | 36 |
| | 4 | 664.6 | 31 | 664.9 | 31 |
| | 8 | 832.8 | 106 | 832.9 | 106 |
| **20** | 5 | 704.1 | 15 | 703.8 | 16 |
| | 6 | 740.5 | 15 | 739.1 | 15 |
| | 7 | 782.8 | 20 | 779.7 | 20 |
| | 8A | 864.7 | 21 | 864.0 | 22 |
| | 11 | 1613.7 | 91 | 1610.4 | 94 |
| | 12 | 2202.4 | 175 | 2185.7 | 185 |
| **60** | 1 | 442.7 | 21 | 442.2 | 21 |
| | 9 | 945.1 | 20 | 943.2 | 20 |
| | 10 | 1373.5 | 31 | 1376.9 | 30 |

**Πίνακας 1.2:** Μήκος κύματος, εύρος και χωρική διακριτκή ικανότητα για κάθε κανάλι των δορυφόρων Sentinel-2A/2B.

Στην εργασία αυτή θα μας απασχολήσουν τα κανάλια 2, 3 και 4 που αντιστοιχούν στα τρία κανάλια του οπτικού φάσματος, μπλε, πράσινο και κόκκινο αντίστοιχα (RGB). Όπως φαίνεται και

[4] Visible and Near-Infrared
[5] Short-Wavelength Infrared

στον Πίνακα 1.2, τα προαναφερθέντα έχουν χωρική διακριτική ικανότητα 10m η οποία ερμηνεύεται ως μια περιοχή $100 \times 100\ km^2$ στην επιφάνεια της Γης.

Η ραδιομετρική ικανότητα του οργάνου MSI είναι 12-bit επιτρέποντας στα πολυφασματικά δεδομένα να λαμβάνουν τιμές έντασης φωτός σε ένα δυναμικό εύρος 0 - 4095. Οι μετρήσεις αυτές μετατρέπονται σε τιμές ανακλαστικότητας και αποθηκεύονται ως 16-bit ακέραιοι αριθμοί.

Τα προϊόντα δεδομένων που προσφέρονται στους χρήστες μέσω της αποστολής Sentinel-2 είναι δύο, τα προϊόντα **επιπέδου 1C** και τα προϊόντα **επιπέδου 2A**. Αμφότερα αφορούν ορθοφωτογραφίες, ενώ τα τελευταία, αποδίδουν τιμές ανακλαστικότητας στο έδαφος, έχουν υποστεί ατμοσφαιρική διόρθωση (σε αντίθεση με τα προϊόντα επιπέδου 1C), μπορούν να χρησιμοποιηθούν άμεσα σε εφαρμογές χωρίς περαιτέρω επεξεργασία (ARD[6]) και είναι αυτά που θα χρησιμοποιηθούν στα πειράματα του αλγορίθμου μας.

[6] Analysis Ready Data

# ΚΕΦΑΛΑΙΟ 2

# Transformers - Axial Attention

Στο κεφάλαιο αυτό θα μιλήσουμε για ορισμένες πιο σύνθετες έννοιες της Βαθιάς Μάθησης, όπως τα μοντέλα αυτό-παλινδρόμησης, τους Transformers και την τεχνική του Self-Attention που χρησιμοποιούν, καθώς και τη προσέγγιση του Axial Attention στην οποία και βασίζεται ο αλγόριθμος μας.

## 2.1 Μοντέλα Αυτό-Παλινδρόμησης

Τα αυτό-παλινδρομικά μοντέλα (AR-Autoregressive models) προβλέπουν μια τιμή της χρονοσειράς $Y$ βασιζόμενα αποκλειστικά στις τιμές του παρελθόντος (lags) δημιουργώντας, μεταξύ αυτών, μια γραμμική σχέση.

Το απλούστερο αυτό-παλινδρομικό μοντέλο (Εξίσωση 2.1) βασίζεται μόνο στην αμέσως προηγούμενη χρονική στιγμή της, υπό εξέτασης, τιμής και χαρακτηρίζεται *πρώτου βαθμού* ή αλλιώς *AR(1)* μοντέλο.

$$Y_t = \omega + \phi \cdot Y_{t-1} + \varepsilon_t \tag{2.1}$$

$$-1 < \phi < 1 \tag{2.2}$$

όπου $\omega$ μια αριθμητική σταθερά, $Y_t$ η τιμή της χρονοσειράς $Y$ την χρονική στιγμή $t$ για την οποία πρόκειται να γίνει η πρόβλεψη, $Y_{t-1}$ η τιμή της ίδιας χρονοσειράς την αμέσως προηγούμενη χρονική στιγμή, $\varepsilon_t$ ένας αριθμητικός παράγοντας που δηλώνει το σφάλμα πρόβλεψης $(Y_t - \hat{Y}_t)$ και $\phi$ μια αριθμητική σταθερά και παράμετρος του μοντέλου, η οποία ερμηνεύει το μέρος της εκάστοτε προηγούμενης στιγμής (στην περίπτωση αυτή της $t-1$) που παρέμεινε αναλλοίωτο στο μέλλον.

Η αναδρομή που προαναφέρθηκε, γίνεται για κάθε τιμή της χρονοσειράς εκτός από την πρώτη. Με τη μέθοδο αυτή, το τελικό αποτέλεσμα επηρεάζεται από όλες τις τιμές των προηγούμενων

χρονικών στιγμών, και τα αντίστοιχα μοντέλα χαρακτηρίζονται ως «μακράς μνήνης», με ακολουθία εξισώσεων:

$$Y_t = \omega + \phi \cdot Y_{t-1} + \varepsilon_t$$

$$Y_{t-1} = \omega + \phi \cdot Y_{t-2} + \varepsilon_{t-1}$$

$$\dots$$

$$Y_t = \frac{\omega}{1-\phi} + \phi^t \cdot Y_1 + \phi^{t-1} \cdot \varepsilon_2 + \phi^{t-2} \cdot \varepsilon_3 + \cdots + \varepsilon_t$$

Διαπιστώνεται πως οι τιμές της χρονοσειράς που απέχουν μεγάλο χρονικό διάστημα από την, υπό εξέταση, τιμή έχουν ολοένα και μικρότερη επίδραση σε αυτήν, εφόσον ισχύει η Εξίσωση 2.2.

Αντίστοιχα δομούνται και τα αυτό-παλινδρομητικά μοντέλα δεύτερου (*AR(2)*) ή μεγαλύτερου βαθμού. Στα AR(2) μοντέλα, η χρονοσειρά είναι μια γραμμική εξίσωση των δύο προηγούμενων χρονικών στιγμών της μορφής:

$$Y_t = \omega + \phi_1 \cdot Y_{t-1} + \phi_2 \cdot Y_{t-2} + \varepsilon_t \tag{2.3}$$

Γενικεύοντας τα παραπάνω, καταλήγουμε στην έννοια των μοντέλων *AR(p)* τα οποία είναι μια γραμμική εξίσωση των p προηγούμενων χρονικών στιγμών:

$$Y_t = \omega + \phi_1 \cdot Y_{t-1} + \phi_2 \cdot Y_{t-2} + \cdots + \phi_p \cdot Y_{t-p} + \varepsilon_t \tag{2.4}$$

Ο βαθμός του μοντέλου που χρησιμοποιείται εξαρτάται από τα δεδομένα και τη φύση της μελέτης. Συνήθως όσο πολυπλοκότερα είναι τα δεδομένα τόσο περισσότερες παράμετροι χρησιμοποιούνται, ενισχύοντας παράλληλα την ποιότητα των προβλέψεων.

## 2.2 Self-Attetnion & Δίκτυα Transformers

### 2.2.1 Self-Attention

Ως μηχανισμός λαμβάνει $n$ πλήθος εισόδων και επιστρέφει $n$ πλήθος εξόδων. Η διεργασία που συμβαίνει κατά τη διάρκεια του, επιτρέπει στις εισόδους να αλληλεπιδράσουν μεταξύ τους και να διαπιστώσουν σε ποιό μέρος της εισερχόμενης πληροφορίας να εστιάσουν περισσότερο και σε ποιό λιγότερο, βάση της συνάφειας από τα συμφραζόμενα. Οι έξοδοι δεν είναι τιμές αντίστοιχες των εισόδων, αλλά αθροίσματα βαρών και «βαθμολογίες προσοχής» για κάθε μια. Ας δούμε όμως, πως λειτουργεί αυτός ο μηχανισμός σε μερικά απλά βήματα:

**Βήμα 1. Προετοιμασία δεδομένων εισόδου και αρχικοποίηση βαρών**: Έστω ότι έχουμε $n$ αριθμό εισόδων διαστάσεων $d$ η κάθε μια, οπότε προκύπτει ο πίνακας $X \in \mathbb{R}^{n \times d}$ και στη συνέχεια αρχικοποιούνται οι πίνακες βαρών $W_Q \in \mathbb{R}^{d \times d_q}$, $W_K \in \mathbb{R}^{d \times d_k}$ και $W_V \in \mathbb{R}^{d \times d_v}$.

**Βήμα 2. Εξαγωγή των διανυσμάτων Q, K, V**: Στο βήμα αυτό πραγματοποιείται ο πολλαπλασιασμός μεταξύ των εισόδων και των αντίστοιχων βαρών όπως φαίνεται στην ομάδα εξισώσεων 2.5.

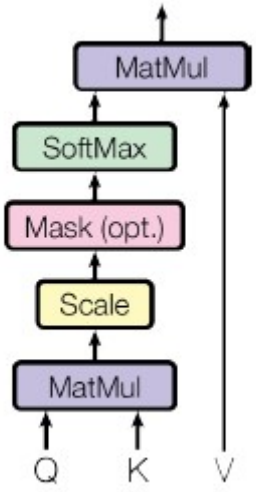
$$Q = X \cdot W_Q \qquad K = X \cdot W_K \qquad V = X \cdot W_V \tag{2.5}$$

όπου $Q \in \mathbb{R}^{n \times d_q}$, $K \in \mathbb{R}^{n \times d_k}$, $V \in \mathbb{R}^{n \times d_v}$ και $d_q = d_k$. Η διάσταση $d_v$ μπορεί να διαφέρει από τις άλλες δύο και ταυτίζεται πάντα με την διάσταση της εξόδου του self-attention.

**Βήμα 3. Υπολογισμός «βαθμολογιών προσοχής»**: Πολλαπλασιάζεται το διάνυσμα $q$ που αντιστοιχεί σε μια είσοδο, με το διάνυσμα $K$ όλων των εισόδων, συμπεριλαμβανομένου και του εαυτού της. Η διαδικασία αυτή ονομάζεται *dot product attention*.

**Βήμα 4. Κανονικοποίηση «βαθμολογιών προσοχής»**: Είναι ένα ενδιάμεσο βήμα, (προτάθηκε από [6]), κατά το οποίο διαιρούνται οι «βαθμολογίες προσοχής» που προκύπτουν, με τη τετραγωνική ρίζα της διάστασης $d_k$. Η διαδικασία αυτή ονομάζεται *scaled dot product attention* (Σχήμα 2.1). Με την αποκλειστική χρήση της προσέγγισης του Βήματος 3, προκύπτουν, ανά περιπτώσεις, μεγάλες τιμές στις «βαθμολογίες προσοχής», οι οποίες οδηγούν σε πολύ μικρές τιμές ύστερα από την υποβολή τους στη συνάρτηση Softmax, φαινόμενο το οποίο αποτρέπεται με την χρήση του παράγοντα $(\frac{1}{\sqrt{d_k}})$.

**Βήμα 5. Εφαρμόζεται η συνάρτηση Softmax** στα αποτελέσματα του προηγούμενου βήματος.




**Σχήμα 2.1:** Scaled Dot-Product Attention.

**Βήμα 6.** Πολλαπλασιάζονται τα αποτελέσματα της συνάρτησης Softmax με τον πίνακα τιμών $V$ και προκύπτουν οι «σταθμισμένες τιμές».

**Βήμα 7.** Τελευταίο βήμα αποτελεί το άθροισμα, ανά στοιχείο, των σταθμισμένων τιμών καταλήγοντας στην τελική έξοδο του attention μηχανισμού που αναφέρεται στην επιλεγμένη είσοδο. Η διαδικασία από το Βήμα 3 και μετά επαναλαμβάνεται για όλες τις εισόδους ξεχωριστά.

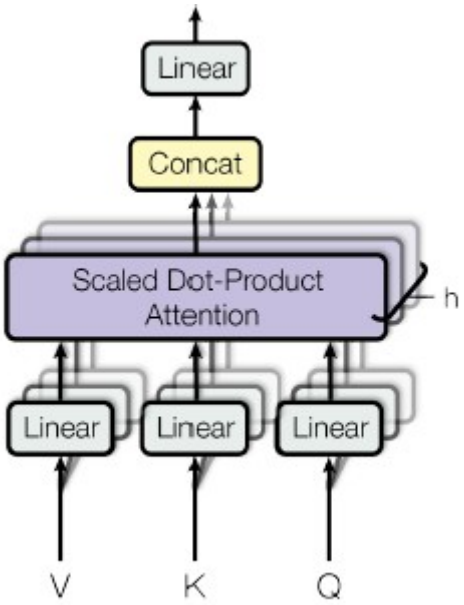
Τα παραπάνω βήματα συνοψίζονται στην εξίσωση 2.6:

$$Attention(Q, K, V) = softmax\left(\frac{K \cdot Q^T}{\sqrt{d_k}}\right) \cdot V \tag{2.6}$$

Προεκτείνοντας τον παραπάνω μηχανισμό έχει προταθεί η μέθοδος του πολυεπίπεδου self-attention (multi-headed) (Σχήμα 2.2). Στην ουσία, η προσέγγιση αυτή, είναι πολύ παρόμοια με την προηγούμενη, με τη διαφοροποίηση ότι τώρα, ορίζονται πολλά επίπεδα self-attention τα οποία λειτουργούν παράλληλα και ανεξάρτητα μεταξύ τους. Ο αριθμός $h$ των επιπέδων που χρησιμοποιούνται κάθε φορά, δεν είναι προκαθορισμένος και αποτελεί υπερπαράμετρο του μοντέλου. Ας δούμε, όμως, βήμα βήμα συνοπτικά πως λειτουργεί:

**Βήμα 1.** Ορίζεται ο αριθμός των επιπέδων $h$ που θα χρησιμοποιηθούν. Κάθε επίπεδο (head) έχει τους δικούς του πίνακες βαρών, οι οποίοι ορίζονται ακριβώς όπως περιγράψαμε προηγουμένως.

**Βήμα 2.** Εξάγονται τα διανύσματα $Q$, $K$, $V$, πολλαπλασιάζοντας τις εισόδους με τους πίνακες βαρών **για κάθε επίπεδο ξεχωριστά** και στη συνέχεια εφαρμόζεται, σε αυτά, η τεχνική του




**Σχήμα 2.2:** Multihead Attention.

self-attention.

**Βήμα 3.** Στο τελευταίο βήμα πραγματοποιείται η συνένωση όλων των εξόδων και πολλαπλασιάζεται με τον πίνακα βαρών $W^O$ με σκοπό την επιθυμητή τελική έξοδο (Εξίσωση 2.7).

$$MultiHead(Q, K, V) = concat(head_1, head_2, \ldots, head_n) \cdot W^O \tag{2.7}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.8}$$

Η μέθοδος του πολυεπίπεδου self-attention διευρύνει την ικανότητα του μοντέλου να εστιάζει σε πλήθος διαφορετικών στοιχείων και θέσεων στα δεδομένα εισόδου και αποτελεί μέρος της αρχιτεκτονικής στην προτεινόμενη προσέγγιση αυτής της εργασίας.

### 2.2.2 Transformers

Το 2017 προτάθηκε ένα καινοτόμο μοντέλο βασισμένο στην τεχνική του πολύ-επίπεδου self-attention με την ονομασία «Transformer». Αξίζει να σημειωθεί ότι μια από τις βασικότερες και πιο χαρακτηριστικές εφαρμογές του ήταν η αυτόματη μετάφραση κειμένων, ωστόσο οι προοπτικές για την χρήση του σε πλήθος άλλων εφαρμογών είναι μεγάλες, κάνοντας φυσικά, τις απαραίτητες τροποποιήσεις.

Εν συντομία θα αναφερθούμε στη αρχιτεκτονική του Transformer, όπως φαίνεται στο Σχήμα 2.3, που χρησιμοποιεί μια δομή Encoder-Decoder, οι οποίοι λειτουργούν παράλληλα. Αμφότε-

ροι μπορούν να δομηθούν με $N$ αριθμό πανομοιότυπων στρώσεων.

**Encoder**

Κάθε στρώση αποτελείται από δύο επίπεδα, το πρώτο εκ των οποίων είναι ο μηχανισμός του πολυεπίπεδου self-attention και το δεύτερο ένα απλό, πλήρως συνδεδεμένο δίκτυο τροφοδοσίας (*fully connected feed-forward network*). Σε κάθε υπόστρωμα εφαρμόζεται μια σχέση παραλλειπτική (*residual connection*), ακολουθούμενη από ένα επίπεδο κανονικοποίησης, αποκτώντας τη μορφή $LayerNorm(x + Sublayer(x))$.

**Decoder**

Διατηρεί τα δύο επίπεδα του Encoder αλλά διαθέτει ένα ακόμα, το οποίο εφαρμόζει πολυεπίπεδο self-attention στις εξόδους του Encoder. Επιπλέον, τροποποιείται το επίπεδο που αφορά το self-attention της προηγούμενης παραγράφου με ένα μασκάρισμα, ώστε να μην «βλέπει» σε μελλοντικές τιμές και η πρόβλεψη μιας θέσης $i$ να εξαρτάται αποκλειστικά από τα αποτελέσματα των θέσεων $< i$.

**Λεπτομέρειες Μοντέλου**

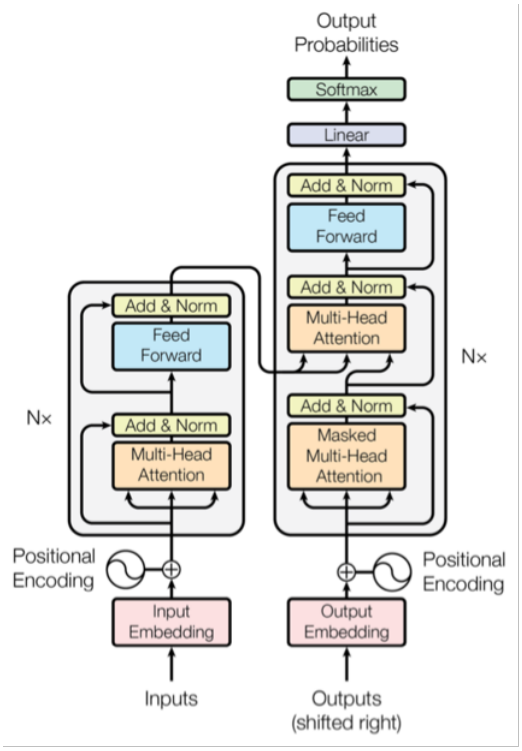
Κάθε στρώση ($N$ σε αριθμό) των παραπάνω δομών περιλαμβάνει ένα πλήρως συνδεδεμένο δίκτυο τροφοδοσίας, το οποίο εφαρμόζεται για κάθε θέση ξεχωριστά και πανομοιότυπα. Το δίκτυο αυτό αποτελείται από δύο γραμμικούς μετασχηματισμούς με μια συνάρτηση ενεργοποίησης *ReLU* ενδιάμεσα.

Χρησιμοποιούνται δύο στρώματα ενσωμάτωσης (Embeddings) με σκοπό να μετατρέψουν τις εισόδους και τις εξόδους σε διανύσματα διαστάσεων $d$. Οι έξοδοι του Decoder υποβάλλονται σε γραμμικό μετασχηματισμό και στη συνέχεια περνούν από μια συνάρτηση Softmax για να ερμηνεύονται, τελικά, ως πιθανότητες για τις προβλεπόμενες τιμές. Τέλος, με σκοπό την αξιοποίηση της φυσικής σειράς των δεδομένων εισόδου, εφαρμόζεται μια κωδικοποίηση θέσεων, γνωστή ως Positional Encoding (PE), η οποία αποθηκεύει πληροφορία για την απόλυτη ή σχετική θέση κάθε στοιχείου.

**Χρήσεις του Attention στον Transformer**

Η παράγραφος αυτή είναι πολύ σημαντική καθώς θα αναλύσουμε πως συνδυάζεται το πολύεπίπεδο self-attention που μελετήσαμε στην Υποενότητα 2.2.1, με την δομή του Transformer μοντέλου σε τρεις διαφορετικές περιπτώσεις:

1. Τα στρώματα του «Encoder-Decoder attention». Εδώ τα queries λαμβάνονται από το προηγούμενο επίπεδο του Decoder, ενώ τα keys, values τροφοδοτούνται από την έξοδο του Encoder. Η αλληλουχία αυτή, επιτρέπει σε κάθε θέση του Decoder να διευρύνει το οπτικό του πεδίο σε όλη την ακολουθία εισόδου.




**Σχήμα 2.3:** Αρχιτεκτονική Μοντέλου Transformer.

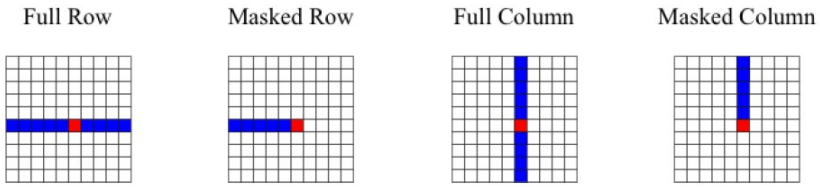
2. Ο Encoder διαθέτει στρώσεις πολύ-επίπεδου self-attention, κατά το οποίο όλα τα q, k, v διανύσματα προέρχονται από το προηγούμενο επίπεδο του. Αποτέλεσμα αυτού, κάθε θέση του Encoder να «κοιτάζει» σε όλες τις θέσεις της προηγούμενης στρώσης του.

3. Παρόμοια λειτουργεί και το self-attention στην δομή του Decoder όπου όλα τα q, k, v διανύσματα προέρχονται, αντίστοιχα, από το προηγούμενο επίπεδο του, οπότε έχει τη δυνατότητα να «κοιτάζει» σε όλες τις προηγούμενες τιμές, συμπεριλαμβανομένης και της τρέχουσας. Υπάρχει μια βασική διαφορά, ώστε να διατηρηθεί η έννοια της αυτό-παλινδρόμησης (βλ. Ενότητα 2.1). Ο Decoder δεν επιτρέπεται να λαμβάνει πληροφορία από μελλοντικές τιμές και για αυτόν τον λόγο μασκάρονται (τιμή $-\infty$) πριν την είσοδό τους στη συνάρτηση Softmax.

## 2.3 Axial Attention

Στην ενότητα αυτή, θα περιγράψουμε μια εναλλακτική δομή Transformer μοντέλου, ένα ακόμα αυτό-παλινδρομητικό μοντέλο κατάλληλο για επεξεργασία πολυδιάστατων δεδομένων, όπως εικόνες ή βίντεο, τον Axial Transformer [10].

Αρχικά ας μελετήσουμε τον θεμέλιο λίθο του μοντέλου, μια γενίκευση της μεθόδου self-attention,




**Σχήμα 2.4:** Από αριστερά: (α) Μη μασκαρισμένο self-attention γραμμής (β) Μασκαρισμένο self-attention γραμμής (γ) Μη μασκαρισμένο self-attention στήλης (δ) Μασκαρισμένο self-attention στήλης. Με μπλε χρώμα συμβολίζεται το οπτικό πεδίο του, υπό εξέταση, κόκκινου εικονοστοιχείου.

η οποία δεν μεταβάλλει τις διαστάσεις των εισόδων και εφαρμόζει attention, μασκαρισμένο η μη, σε επίπεδο αξόνων, ως εκ τούτου και το όνομα Axial attention. Τα δεδομένα μας είναι πολυφασματικές εικόνες οπότε η τεχνική θα περιγραφεί με αυτό ως προϋπόθεση. Ορίζουμε ως $Attention_1$ τη μέθοδο του self-attention όταν εφαρμόζεται σε μια στήλη, διατηρώντας ανεξάρτητες όλες τις υπόλοιπες στήλες και ως $Attention_2$ όταν εφαρμόζεται ανά γραμμή με τον ίδιο τρόπο. Η παραλλαγή αυτή, έρχεται να αντισταθμίσει το υψηλό υπολογιστικό κόστος των πρότυπων δικτύων Transformers, το οποίο για μια εικόνα μεγέθους $N = S \times S$ θα ήταν $\mathcal{O}(N^2)$. Στην περίπτωση του Axial attention το υπολογιστικό κόστος, για την ίδια εικόνα, είναι $\mathcal{O}(N\sqrt{N})$, έχοντας πετύχει εξοικονόμηση βαθμού $\mathcal{O}(\sqrt{N})$. Γενικεύοντας τα παραπάνω για έναν πολυδιάστατο τένσορα $N = S^d$ η μέθοδος αυτή, περιορίζει το κόστος κατά $\mathcal{O}(N^{(d-1)/d})$. Φυσικά η εξοικονόμηση των υπολογιστικών πόρων που αναφέραμε δεν έρχεται χωρίς αντίτιμο και αυτό είναι πως το οπτικό πεδίο του Axial attention περιορίζεται σε έναν μόνο άξονα τη φορά. Στη συνέχεια όμως, θα δούμε με ποιόν τρόπο επιλύεται αυτό το πρόβλημα και το οπτικό πεδίο διευρύνεται σε όλο το μήκος των δεδομένων εισόδου.

Οι χρήσεις της κανονικοποίησης και των συνελίξεων παραμένουν αναλλοίωτες όπως τις είδαμε στην Ενότητα 2.2.2, ενώ εισάγεται η έννοια του $MaskedAttention_k$, μια παραλλαγή της κλασσικής χρήσης της μεθόδου, κατά την οποία το αποτέλεσμα του $MaskedAttention_k(x)$ της θέσης $i$ εξαρτάται αποκλειστικά από τα στοιχεία $1, \ldots, i$ του τένσορα $x$ στον άξονα $k$. Οι τύποι των διάφορων στρώσεων του Axial attention με τα αντίστοιχα οπτικά πεδία τους παρουσιάζονται στο Σχήμα 2.4.

Αφού εξηγήσαμε τη βασική τεχνική, ήρθε η ώρα να ολοκληρώσουμε το κεφάλαιο με την περιγραφή της αρχιτεκτονικής του Axial Transformer, ενός πολυεπίπεδου αυτό-παλινδρομητικού μοντέλου της μορφής $p_\theta(x) = \prod_{i=1}^{N} p_\theta(x_i|x_{<i})$. Τα μπλοκ που θα χρησιμοποιηθούν συντίθενται ως εξής:

- $FFBlock(x) = x + Dense_D(ReLU(Dense_{D'}(LayerNorm(x))))$
- $AttentionBlock_k(x) = x + Dense_D(Attention_k(LayerNorm(x)))$[1]

[1] $k = 1$ για attention στήλης και $k = 2$ για attention γραμμής

- $MaskedAttentionBlock_k(x) = x + Dense_D(MaskedAttention_k(LayerNorm(x)))$
- $TransformerBlock_k(x) = FFBlock(AttentionBlock_k(x))$
- $MaskedTransformerBlock_k(x) = FFBlock(MaskedAttentionBlock_k(x))$

Για εικόνα $x$ ενός καναλιού οι διαστάσεις είναι $H \times W$, με κάθε εικονοστοιχείο να λαμβάνει τιμές εντάσεων $[0, 255]$. Το κανάλι αυτό αρχικά μετατρέπεται σε έναν τένσορα $h$, διαστάσεων $H \times W \times D$. Σκοπός του μοντέλου είναι ο μετασχηματισμός του $h$, σε έναν νέο τένσορα $H \times W \times 256$ ο οποίος θα αντιπροσωπεύει πιθανότητες για κάθε τιμή των εικονοστοιχείων, υπό την προϋπόθεση οι τιμές αυτές να προκύπτουν μόνο από προηγούμενα εικονοστοιχεία του $x$ σύμφωνα με την σειρά σάρωσης (το σημείο (0,0) ταυτίζεται με την πάνω αριστέρα γωνία). Για τον σκοπό αυτό, το μοντέλο απαρτίζεται από δύο αλληλοεξαρτώμενους κλάδους:

**Outer Decoder:** Επεξεργάζεται πληροφορία για τις τιμές όλων των προηγούμενων γραμμών $x_{<i,:}$ χρησιμοποιώντας $N$ αριθμό, *μη μασκαρισμένων self-attention γραμμής* και *μασκαρισμένων self-attention στήλης*, επιπέδων. Η πληροφορία αυτή μετατοπίζεται μια γραμμή προς τα κάτω ώστε η έξοδος του Outer Decoder να μην επηρεάζεται από την τρέχουσα σειρά $i$.

$$h = Embeddings(x) \tag{2.9}$$

$$u = h + PositionalEmbeddings \tag{2.10}$$

$$u = MaskedTransformerBlock_1(TransformerBlock_2(u)) \times N \tag{2.11}$$

$$c^O = ShiftDown(u) \tag{2.12}$$

**Inner Decoder:** Σε συνέχεια του Outer Decoder κατά τον οποίο έχει ληφθεί πληροφορία για όλες τις προηγούμενες γραμμές, σειρά έχουν τα εικονοστοιχεία της ίδιας γραμμής αλλά των προηγούμενων στηλών $x_{i,<j}$. Οι είσοδοι $h$ του μοντέλου μετατοπίζονται μια θέση δεξιά, ώστε να μασκαριστεί το τρέχων εικονοστοιχείο ενώ παράλληλα λαμβάνουν πληροφορία από την έξοδο του Outer Decoder. Εφαρμόζονται $N$ επίπεδα μασκαρισμένων self-attention γραμμής οπότε το τέλικο αποτέλεσμα $c^I_{i,j}$ συντίθεται με δεδομένα από κάθε εικονοστοιχείο $x_{<i,:}$ και $x_{i,<j}$. Με τη μέθοδο αυτή επιλύεται το πρόβλημα του περιορισμένου οπτικού πεδίου που είχαμε προαναφέρει και το μοντέλο αποκτά, έμμεσα, πλήρες οπτικό πεδίο όλων των δεδομένων.

Η έξοδος του Inner Decoder $c^I$ περνά, σε τελικό στάδιο, από ένα επίπεδο κανονικοποίησης και μια συνέλιξη ώστε να προκύψει το επιθυμητό αποτέλεσμα, δηλαδή ένας τένσορας πιθανοτήτων $H \times W \times 256$.

$$h = c^O + ShiftRight(h) + PositionalEmbeddings \tag{2.13}$$

$$c^I = MaskedTransformerBlock_2(h) \times N \tag{2.14}$$

$$p(x_{i,j}) = Dense(LayerNorm(c^I)) \tag{2.15}$$

Στην περίπτωση που η είσοδος είναι μια εικόνα διαστάσεων $H \times W \times C$ όπου $C$ το πλήθος των καναλιών της, τότε το μοντέλο λαμβάνει πληροφορία από τις τιμές όλων των προηγούμενων καναλιών από τον **Encoder**. Κάθε κανάλι περνά από $N$ επίπεδα μη μασκαρισμένων self-attention γραμμής και στήλης. Τα αποτελέσματά τους αθροίζονται και συγκεντρώνονται σε μια έξοδο $c$ η οποία δίνει πληροφορία στον Outer και Inner Decoder μέσω των εξισώσεων 2.9, 2.13 αντίστοιχα.

# ΚΕΦΑΛΑΙΟ 3

# Σχετικές Εργασίες στη Βιβλιογραφία

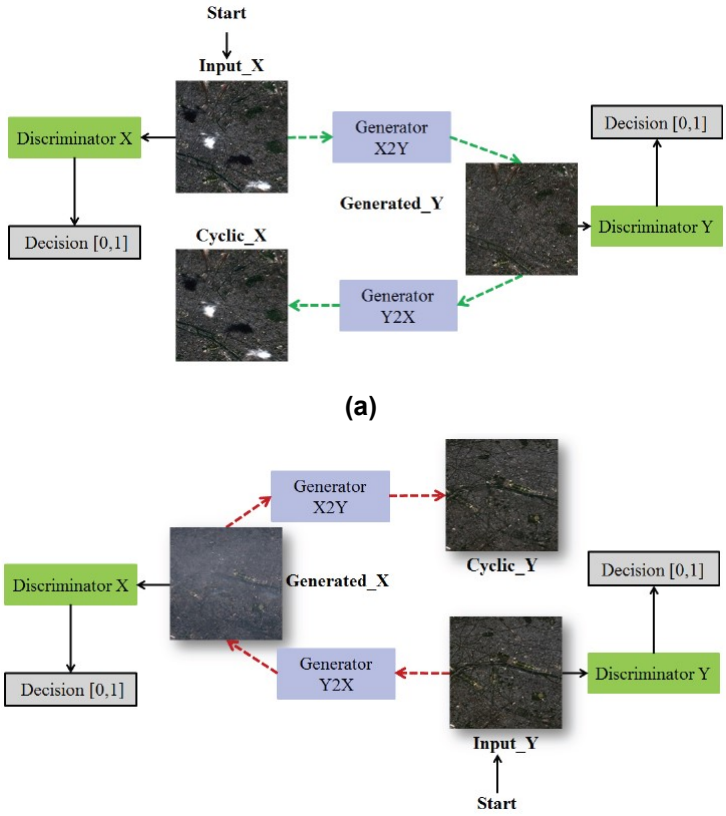
Στο κεφάλαιο αυτό θα αναλύσουμε εν συντομία μερικές από τις υφιστάμενες τεχνικές που έχουν αναπτυχθεί, με σκοπό την αφαίρεση συννέφων από πολυφασματικές δορυφορικές εικόνες. Οι προσεγγίσεις αυτές, διαχωρίζονται σε δύο μεγάλες ενότητες με κριτήριο, την χρήση η μη, πολυχρονικών δεδομένων.

## 3.1 Τεχνικές βασισμένες σε μια εικόνα

### 3.1.1 Cloud-GAN

Η προσέγγιση αυτή [7] χρησιμοποιεί την μέθοδο των **GANs** (Generative Adversarial Networks) [2] για την επίλυση του ζητήματος, ένα σύμπλεγμα ανταγωνιστικών, μεταξύ τους, δικτύων που αποτελούνται συνήθως από έναν Generator και έναν Discriminator. Προτάθηκαν για πρώτη φορά το 2014 και η επιτυχία τους σε εφαρμογές όπως η επεξεργασία ή παραγωγή νέων εικόνων και το γέμισμα κενών, έγκειται στην ανταγωνιστική συνάρτηση κόστους (adversarial loss) που εισήγαγαν. Με τη χρήση της συνάρτησης αυτής, τα GANs δύναται να εκπαιδευτούν ώστε να παράγουν εξ ολοκλήρου νέες «ψεύτικες» εικόνες η οποίες «μοιάζουν» (προέρχονται δηλαδή από την ίδια κατανομή) με τις αυθεντικές. Στην πραγματικότητα, το καθήκον του Discriminator είναι η αναγνώριση μεταξύ αυθεντικών και παραγόμενων εικόνων, ενώ ο Generator αναλαμβάνει την παραγωγή των νέων εικόνων με σκοπό οι τελευταίες να μην γίνουν αντιληπτές από τον Discriminator.

Η τεχνική του Cloud-GAN (2018) διαχωρίζει τα δεδομένα σε δύο κλάδους, $X$ για τις εικόνες επηρεασμένες από σύννεφα και $Y$ για τις καθαρές εικόνες. Εισάγει δύο συναρτήσεις $G : X \to Y$ και $F : Y \to X$ οι οποίες μοντελοποιούνται με τη βοήθεια δύο δικτύων, του $Generator_{X2Y}$ και του $Generator_{Y2X}$ αντίστοιχα (Σχήμα 3.1). Ορίζονται δύο δίκτυα Discriminator, $D_X$ και $D_Y$ που καλούνται να διαχωρίσουν τα πραγματικά δεδομένα $(x, y)$ από τα παραγόμενα $F(Y)$ και $G(X)$. Ο τελικός στόχος του μοντέλου περιγράφεται ως εξής:




**Σχήμα 3.1:** 3.1a: Συνοχή προς τα εμπρός $x \to G(x) \to F(G(x)) \approx x$, 3.1b: Συνοχή προς τα πίσω $y \to F(y) \to G(F(y)) \approx y$.

$$\min_{G,F} \mathcal{L}_{Gen}(G,F) = \mathcal{L}_{LSGAN}(G,X,Y) + \mathcal{L}_{LSGAN}(F,Y,X) + \mathcal{L}_{cyc}(G,F) \tag{3.1}$$

όπου $\mathcal{L}_{LSGAN}$ η συνάρτηση κόστους, όπως ορίζεται στα ομώνυμα GANs ελαχίστων τετραγώνων, για τα δύο δίκτυα $G, F$. Η καινοτομία της τεχνικής έγκειται στον τρίτο παράγοντα της Εξίσωσης 3.1. Ο όρος $\mathcal{L}_{cyc}(G,F)$ περιγράφει μια συνάρτηση κόστους, με σκοπό τη διατήρηση της κυκλικής συνοχής του μοντέλου.

Πλεονέκτημα του μοντέλου αποτελεί το γεγονός ότι κατά την εκπαίδευση, δεν χρειάζεται ζεύγη εικόνων επηρεασμένων από σύννεφα και των αντίστοιχων καθαρών τους, ούτε ειδικές ζώνες δεδομένων όπως κανάλια εγγύς υπέρυθρου ή SAR που επιτυγχάνουν μια μερική διείσδυση στα νέφη. Στον αντίποδα, το μοντέλο δεν αποδίδει σωστά με σύννεφα υψηλών πυκνοτήτων.

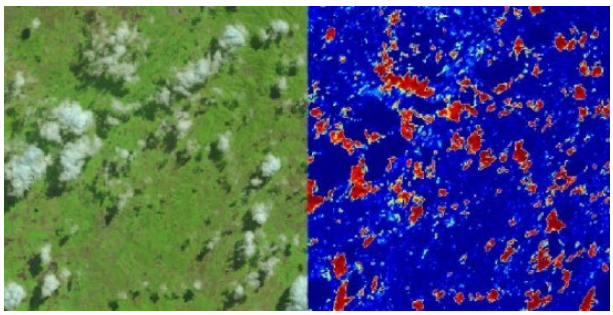
### 3.1.2 SpA-GAN

Χρησιμοποιεί την μέθοδο των GANs με ένα δίκτυο Generator ($G$) και ένα δίκτυο Discriminator ($D$). Το SpA-GAN (2020) [13] δέχεται ως δεδομένα εκπαίδευσης ζεύγη, καθαρών και επηρεασμένων από σύννεφα, πολυφασματικών εικόνων RGB της ίδιας περιοχής. Ο Generator είναι ένα συνελικτικό νευρωνικό δίκτυο που ονομάζεται SPANet (Spatial Attentive Network) και δομείται κατά σειρά από μια συνέλιξη των δεδομένων ώστε να εξαχθεί η αρχική πληροφορία των στοιχείων, τρεις κλασσικές στρώσεις παραλλειπτικών διεργασιών (residual blocks), τέσσερα μπλοκ «χωρικής προσοχής» (SAB), δύο ακόμα στρώσεις παραλλειπτικών διεργασιών και από μια τελική συνέλιξη εκ της οποίας προκύπτει η παραγόμενη εικόνα απαλλαγμένη από σύννεφα. Τα μπλοκ SAB χρησιμοποιούνται για τον υπολογισμό της προσοχής που απαιτείται να δωθεί σε κάθε εικονοστοιχείο, αναλύοντας τη χωρική κατανομή των συννέφων. Στο Σχήμα 3.2 παρουσιάζεται οπτικά, σε παλέττα χρωματικών θερμοτήτων, το αποτέλεσμα της χωρικής προσοχής. Ο Discriminator είναι ένα συνελικτικό νευρωνικό δίκτυο που αποτελείται από έξι στρώσεις συνελίξεων και ανάμεσα τους πέντε επίπεδα κανονικοποίησης και συναρτήσεων ενεργοποίησης Leaky ReLU.

Η εξίσωση της συνάρτησης κόστους για το SpA-GAN μοντέλο ορίζεται με τρεις βασικούς παράγοντες:

$$\mathcal{L}_{SpA} = \arg\min_{G}\max_{D} \mathcal{L}_{cGAN}(G, D) + \mathcal{L}_{L1}(G) + \mathcal{L}_{Att} \tag{3.2}$$

όπου $\mathcal{L}_{cGAN}$ η συνάρτηση κόστους, όπως τυπικά ορίζεται στα δίκτυα GANs [5] , $\mathcal{L}_{L1}$ η συνάρτηση κόστους L1 που χρησιμοποείται για τον υπολογισμό της ακρίβειας κάθε αναδομημένου εικονοστοιχείου και $\mathcal{L}_{Att}$ η συνάρτηση κόστους των επιπέδων «χωρικής προσοχής».



**Σχήμα 3.2:** Θερμικός χάρτης χωρικής προσοχής, προϊόν του SPANet.

## 3.2 Τεχνικές βασισμένες σε χρονοσειρές

Οι τεχνικές που παρουσιάστηκαν ενδεικτικά στην προηγούμενη ενότητα και όσες ανήκουν στην ίδια κατηγορία με αυτές, μειονεκτούν στο γεγονός πως δεν εκμεταλλεύονται τον άξονα του χρόνου στα δεδομένα, οπότε δεν υπάρχει καμία ιστορική πληροφορία για το εκάστοτε εικονοστοιχείο. Αντιθέτως σε αυτή την ενότητα, θα αναλύσουμε δύο μεθόδους που χρησιμοποιούν τον άξονα του χρόνου, έχοντας ως δεδομένα είτε βίντεο (καρέ) ή πολυφασματικές εικόνες (κύβους χρονοσειρών).

### 3.2.1 Learnable Gated Temporal Shift Module

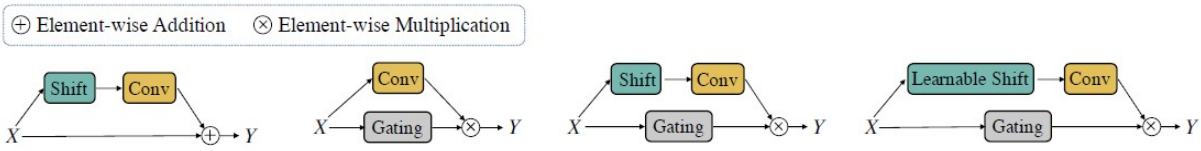
Το LGTSM (2019) [9] επιχειρεί να γεμίσει τα κενά μασκαρισμένων καρέ από βίντεο, μοντελοποιώντας τα χωρικά και χρονικά χαρακτηριστικά των δεδομένων. Αποφεύγει την κοστοβόρα χρήση των τρισδιάστατων συνελίξεων και δομείται βασισμένο σε δισδιάστατες συνελίξεις που εκμεταλλεύονται τα γειτονικά καρέ.

Θεμέλιο λίθο του μοντέλου, αποτελεί η χρονικά μετατοπισμένη μονάδα, **TSM**, η οποία για κάθε συνελικτική στρώση δέχεται δεδομένα διαστάσεων $(B, C, L, H, W)$ όπου $B$ το πλήθος των δεδομένων προς ταυτόχρονη επεξεργασία, $C$ ο αριθμός των καναλιών, $L$ η συνολική χρονική διάρκεια, και $H, W$ το ύψος και το πλάτος των δεδομένων αντίστοιχα. Για κάθε καρέ της χρονικής διάρκειας $L$, η μονάδα TSM μετατοπίζει ένα τμήμα καναλιών στο προήγουμενο και επόμενο καρέ προτού πραγματοποιηθούν οι διασδιάστατες συνελίξεις, με αποτέλεσμα τα κανάλια αυτά, να περιέχουν πληροφορία από πολλαπλές χρονικές στιγμές. Το πρόβλημα έγκειται στο γεγονός ότι η TSM μονάδα δεν είναι ικανή να διαχωρίσει αν τα στοιχεία που δέχεται είναι έγκυρα ή μη, δηλαδή αν αναφέρονται σε μασκαρισμένες περιοχές δίχως πληροφορία. Λύση στο ζήτημα αυτό δίνει η εφαρμογή ενός συνελικτικού φίλτρου (Gating), συνθέτοντας μια νέα μονάδα, την **GTSM** (Gated Temporal Shift Module). Το φίλτρο αυτό, εφαρμόζεται στα δεδομένα εισόδου αλλά και στην έξοδο του μοντέλου λειτουργώντας ως ρυθμιστής εγκυρότητας, ώστε να διαχωρίζεται η φύση της κάθε περιοχής (με μάσκα, χωρίς μάσκα ή «γεμισμένη» από το μοντέλο). Ως τώρα, η GTSM μονάδα λαμβάνει πληροφορίες για κάθε καρέ, αποκλειστικά από τον προηγούμενο και τον επόμενο γείτονα του. Με τη χρήση των εκπαιδεύσιμων, χρονικά μετατοπισμένων μονάδων **LGTSM**, το μοντέλο μπορεί πλέον να λαμβάνει στοιχεία και από μακρινά καρέ.

Η συνάρτηση κόστους σε αυτή την περίπτωση αποτελείται από τέσσερις όρους:

$$\mathcal{L}_{total} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} + \lambda_{G}\mathcal{L}_{G} \tag{3.3}$$

όπου $\mathcal{L}_{L1}$ η γνωστή συνάρτηση κόστους $L1$ που εφαρμόζεται σε κλίμακα εικονοστοιχείου (low-level), $\mathcal{L}_{perc}$ & $\mathcal{L}_{style}$ συναρτήσεις κόστους ώστε να διατηρηθούν ευδιάκριτα τα στοιχεία των




**Σχήμα 3.3:** Από αριστερά: (α) πρότυπη χρονικά μετατοπισμένη μονάδα TSM, (β) συνελικτικό φίλτρο (Gating), (γ) φιλτραρισμένη μονάδα GTSM, (δ) τελικό μοντέλο LGTSM.

εικόνων και $\mathcal{L}_G$ η συνάρτηση κόστους που αναλαμβάνει την ύπαρξη χρονικής συνέπειας και διατηρεί τα αποτελέσματα ρεαλιστικά (high-level). Η τελευταία προκύπτει από ένα ανταγωνιστικό δίκτυο GAN, όπου αμφότεροι Generator και Discriminator διαθέτουν μονάδες GTSM.

### 3.2.2 Fill and Fit

Η FF προσέγγιση εκτελείται σε δύο βήματα, πρώτα «γεμίζει» τα κενά της εκάστοτε χρονοσειράς (*fill*) και ύστερα ένα αρμονικό μοντέλο, είτε γραμμικό ή μη-γραμμικό, εφαρμόζεται στη συμπληρωμένη χρονοσειρά του πρώτου βήματος (*fit*).

Το γέμισμα των κενών αναλαμβάνουν, συνήθως, οι αλγόριθμοι εναλλακτικού παρόμοιου εικονοστοιχείου, ASP[1]. Συμπληρώνουν τις ελλιπείς παρατηρήσεις χρησιμοποιώντας ένα ή περισσότερα παρόμοια εικονοστοιχεία, της ίδιας ή χρονικά κοντινής εικόνας, τα οποία διαθέτουν έγκυρη πληροφορία. Η επιλογή των κατάλληλων τιμών γίνεται βάση δεικτών ομοιότητας. Η προσέγγιση FF χρησιμοποιεί τον αλγόριθμο του, χωρικά και χρονικά, εναλλακτικού παρόμοιου εικονοστοιχείου **SAMSTS** [8], λόγω της ικανότητας του να γεμίζει εκτενή κενά και τη σταθερότητα που παρουσιάζει στις αλλαγές χρήσεων γης, κατά μήκος των χρονοσειρών. Στον SAMSTS αλγόριθμο οι θέσεις των παρόμοιων εικονοστοιχείων εντοπίζονται συγκρίνοντας τις φασματικές υπογραφές των χρονοσειρών, για κάθε κενό εικονοστοιχείο με όσα διαθέτουν πληροφορία, χρησιμοποιώντας την μετρική SAM[2] προσαρμοσμένη στη μερική έλλειψη παρατηρήσεων στον άξονα του χρόνου. Η παραπάνω σύγκριση γίνεται ανά ζώνες και όχι σε ολόκληρη την εικόνα, καθώς το υπολογιστικό κόστος θα ήταν πολύ μεγάλο. Οι ζώνες αυτές, προκύπτουν από την ομαδοποίηση γειτονικών εικονοστοιχείων στον άξονα του χώρου και του χρόνου.

΄Υστερα από την επιλογή των τιμών γεμίσματος για τις κενές παρατηρήσεις, ακολουθεί το δεύτερο βήμα που αφορά τη προσαρμογή τους στο σύνολο της εικόνας. Το έργο αυτό αναλαμβάνουν γραμμικά και μη-γραμμικά αρμονικά μοντέλα. Πειραματικά, με δεδομένα Landsat, έχει αποδειχθεί πως αν υπάρχουν 15-20 έγκυρες παρατηρήσεις, το μη-γραμμικό μοντέλο των 5 παραμέτρων προσαρμόζει αξιόπιστα τις χρονοσειρές, ενώ για $21+$ έγκυρες παρατηρήσεις, προτιμάται το **γραμμικό μοντέλο** της Εξίσωσης 3.4 **με 7 παραμέτρους** [14].

[1] Alternative Similar Pixel
[2] Spectral-Angle-Mapper (μοντελοποίηση φασματικής γωνίας)

$$\hat{\rho}_\lambda(t) = a_{0,\lambda} + \sum_{m=1}^{M} \left( a_{m,\lambda} \cos \frac{2\pi t}{L} + b_{m,\lambda} \sin \frac{2\pi t}{L} \right) \tag{3.4}$$

όπου $\hat{\rho}_\lambda(t)$ η τιμή που προκύπτει από την πρόβλεψη του μοντέλου για τη φασματική ζώνη $\lambda$, τη χρονική στιγμή $t$, $a_{0,\lambda}$ ο συντελεστής Fourier που ορίζεται ως μέση τιμή του $\hat{\rho}_\lambda(t)$ στη χρονοσειρά, $a_{m,\lambda}$ και $b_{m,\lambda}$ οι συντελεστές Fourier κατά την $m$ συνιστώσα, $M = 3$ για μοντέλο 7 παραμέτρων και $L$ η χρονική περίοδος των δεδομένων. Το μοντέλο της Εξίσωσης 3.4 επιλύεται με τη χρήση της μεθόδου Ελαχίστων Τετραγώνων.

### 3.2.3 Τεχνικές βασισμένες στο σετ δεδομένων SEN12MS-CR-TS

Τα χαρακτηριστικά του σετ δεδομένων SEN12MS-CR-TS αναλύονται στην υποενότητα 5.1.2, ωστόσο παράλληλα με τη δημοσίευσή του, προτείνονται δύο νέοι μέθοδοι αφαίρεσης συννέφων [19]. Η πρώτη μέθοδος επεξεργάζεται χρονοσειρές SAR απεικονίσεων και πολυφασματικών δεδομένων από τις αποστολές Sentinel-1 και Sentinel-2 του προγράμματος Copernicus, αντίστοιχα. Σκοπός της είναι η παραγωγή μιας νέας Sentinel-2 απεικόνισης, απαλλαγμένης από σύννεφα. Η αρχιτεκτονική της μεθόδου αποτελείται από ένα βαθύ νευρωνικό δίκτυο βασισμένο στο δίκτυο του Generator του [15]. Ξεκινά με την επεξεργασία και αφαίρεση των συννέφων κάθε χρονικής στιγμής ξεχωριστά με χρήση παραλλειπτικών κλάδων [12] και στη συνέχεια ενοποιεί την παραγόμενη πληροφορία στον άξονα του χρόνου. Τέλος, εφαρμόζονται τρισδιάστατες συνελίξεις (3D Convolutions) για τη σωστή ερμηνεία της χρονικής συνιστώσας.

Η συνάρτηση κόστους προς ελαχιστοποίηση κατά τη διάρκεια εκπαίδευσης του μοντέλου υπολογίζεται από τον τύπο:

$$\mathcal{L}_{all} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{perc}\mathcal{L}_{perc} \tag{3.5}$$

όπου $\lambda_{L1} = 100$ και $\lambda_{perc} = 1$. Η συνάρτηση $\mathcal{L}_{perc}$, σε αυτή την μέθοδο, υπολογίζεται βάση του δικτύου VGG16 [3] προεκπαιδευμένο στο σετ δεδομένων SEN12MS [11] με σκοπό την ταξινόμηση χρήσεων γης.

Η δεύτερη μέθοδος δέχεται για εισόδους αποκλειστικά χρονοσειρές δεδομένων SAR με σκοπό την παραγωγή της αντίστοιχης χρονοσειράς πολυφασματικών απεικονίσεων Sentinel-2, απαλλαγμένης από σύννεφα. Η αρχιτεκτονική της είναι σχετικά απλή, ακολουθώντας ένα τρισδιάστατο μοντέλο Encoder-Decoder κατανεμημένο σε μορφή U-Net [4] με την εφαρμογή παραλλειπτικών συνδέσεων ανά ζεύγη στρώσεων.

Η συνάρτηση κόστους της δεύτερης μεθόδου υπολογίζεται από τον τύπο:

$$\mathcal{L}_{all} = \lambda_{L2}\mathcal{L}_{L2} + \lambda_{perc}\mathcal{L}_{perc} \tag{3.6}$$

όπου $\lambda_{L2} = 1$ και $\lambda_{perc} = 0.01$.

Η χρήση δεδομένων SAR στις προαναφερθείσες μεθόδους καθώς και η φύση των δεδομένων Sentinel-2 (Επίπεδο 1C αντί για Επίπεδο 2A) δεν τις καθιστούν άμεσα συγκρίσιμες με τη δική μας αρχιτεκτονική. Ωστόσο όπως θα δούμε παρακάτω πραγματοποιούνται πειράματα με το προτεινόμενο σετ δεδομένων ώστε να δοκιμαστεί και να επικυρωθεί η εγκυρότητα του μοντέλου μας.

# ΚΕΦΑΛΑΙΟ 4

# Προτεινόμενη Μεθοδολογία - CloudTran

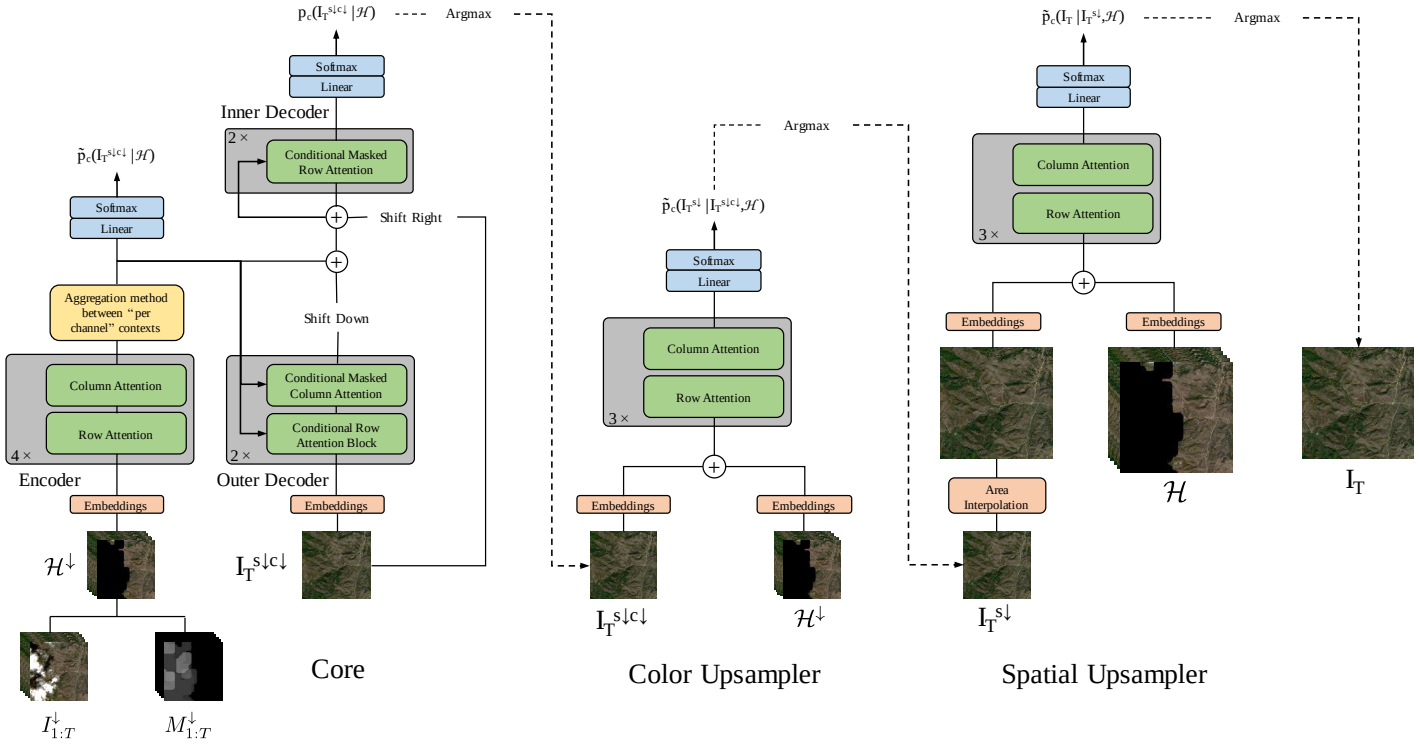
Προτείνουμε μια νέα μέθοδο αφαίρεσης συννέφων από πολυφασματικά δεδομένα, έχοντας ως βασικό εργαλείο την τεχνική του Axial Attention που μελετήσαμε στην Ενότητα 2.3. Η προσέγγιση **CloudTran** [18] ανήκει στην κατηγορία των μοντέλων που αξιοποιούν χρονοσειρές, για την αναδόμηση της νέας απεικόνισης, μελετώντας την ιστορικότητα κάθε εικονοστοιχείου. Με σκοπό τη διαμόρφωση του τελικού, βέλτιστου αλγορίθμου, η αρχιτεκτονική μελετήθηκε σε δύο στάδια. Εμπνευσμένοι από το πιθανολογικό μοντέλο του Colorization Transformer [17], κατά το **πρώτο στάδιο** διαχωρίζουμε το βασικό ζήτημα σε τρία υπο-προβλήματα:

1. Εκτίμηση της νέας εικόνας σε χαμηλή διακριτική ικανότητα και σε χαμηλό φάσμα χρωμάτων.
2. Εκ νέου αναδόμηση της εικόνας στο πλήρες φάσμα χρωμάτων.
3. Τελική αναδόμηση της εικόνας σε υψηλή διακριτική ικανότητα.

Τα υπο-προβλήματα αυτά αναλαμβάνουν να φέρουν σε πέρας τρία ξεχωριστά, ανεξάρτητα μεταξύ τους, δίκτυα, ο «core» (βασικό δίκτυο ανακατασκευής), ο «color upsampler» και ο «spatial upsampler» αντίστοιχα (Σχήμα 4.1). Ωστόσο, κατά το **δεύτερο στάδιο** μελέτης, τα πρώτα δύο υπο-προβλήματα συγχωνεύονται και αντιμετωπίζονται, ταυτόχρονα πλέον, από το βασικό δίκτυο ώστε να αποφευχθεί η χρήση του «color upsampler» (Σχήμα 4.2). Όλα τα προαναφερθέντα δίκτυα αποτελούνται από στρώσεις Axial Attention, το οποίο περιορίζει το υπολογιστικό κόστος σε $\mathcal{O}(N\sqrt{N})$ αντί για $\mathcal{O}(N^2)$ όπου $N$ το μέγεθος της εικόνας και με δύο μόλις επίπεδα (γραμμής και στήλης) αποκτά πλήρες οπτικό πεδίο των δεδομένων μιας απεικόνισης.

Στόχος της μεθόδου είναι η δημιουργία μιας αναδομημένης εικόνας $I_T \in \mathbb{R}^{H\times W\times B}$, απαλλαγμένης από σύννεφα, βάση του κύβου $\mathcal{H} \in \mathbb{R}^{H\times W\times B\times T}$ όπου $H \times W$ οι διαστάσεις των εικόνων, $B$ ο αριθμός των φασματικών καναλιών και $T$ το πλήθος των ιστορικών απεικονίσεων $I_t$ για την εκάστοτε χρονοσειρά. Για κάθε κύβο $\mathcal{H}$ μοντελοποιείται μια κατανομή $p(I_T|\mathcal{H})$, βάση της οποίας προκύπτει η απεικόνιση $I_T$.

Η τελική εικόνα $I_T$, δίχως την επιρροή συννέφων, επιλέγεται να μην αναδομηθεί απευθείας από τον τον κύβο $\mathcal{H}$, αλλά πρώτα να αναδομηθούν οι ενδιάμεσες απεικονίσεις $I_T^{s\downarrow}$ και $I_T^{s\downarrow c\downarrow}$, μειωμέ-



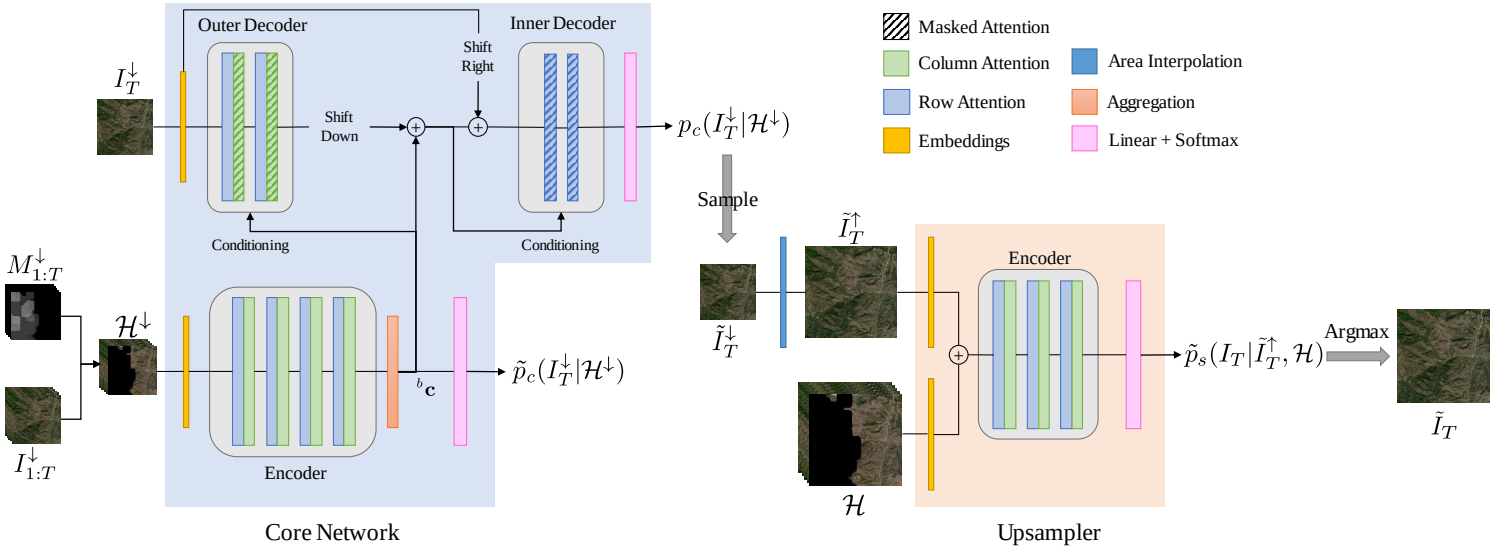

**Σχήμα 4.1:** Πρώτο στάδιο μελέτης της αρχιτεκτονικής του μοντέλου CloudTran, με **τρία** δίκτυα, βασισμένη στην αρχιτεκτονική του Colorization Transformer. (καλύτερα ορατή σε μεγέθυνση)

νης χωρικής και ραδιομετρικής διακριτικής ικανότητας. Διακρίνουμε τα δύο στάδια μελέτης της CloudTran αρχιτεκτονικής:

1. Στην πρώτη περίπτωση, χρησιμοποιούνται και τα τρία δίκτυα, όπως προτείνεται στον Colorization Transformer (Εξίσωση 4.1). Η απεικόνιση $I_T^{s\downarrow} \in \mathbb{R}^{H^\downarrow \times W^\downarrow \times B}$ προκύπτει με χωρική παρεμβολή από την $I_T$ και έχει διαστάσεις $H^\downarrow \times W^\downarrow$ μειωμένης χωρικής διακριτικής ικανότητας. Η απεικόνιση $I_T^{s\downarrow c\downarrow}$ είναι αναπαράσταση της $I_T^{s\downarrow}$ με ραδιομετρική ικανότητα 3-bits, δηλαδή με $2^3 = 8$ εντάσεις ανά κανάλι. Για κάθε εικονοστοιχείο, λοιπόν, μιας RGB εικόνας, ο αριθμός των πιθανών χρωμάτων είναι $8^3 = 512$.

$$\begin{aligned} p(I_T|\mathcal{H}) &= p(I_T|\mathcal{H}) \cdot p(I_T^{s\downarrow c\downarrow}, I_T^{s\downarrow}|I_T, \mathcal{H}) \\ &= p(I_T^{s\downarrow c\downarrow}, I_T^{s\downarrow}, I_T|\mathcal{H}) \\ &= p(I_T|I_T^{s\downarrow}, \mathcal{H}) \cdot p(I_T^{s\downarrow}|I_T^{s\downarrow c\downarrow}, \mathcal{H}^\downarrow) \cdot p(I_T^{s\downarrow c\downarrow}|\mathcal{H}^\downarrow) \end{aligned} \tag{4.1}$$

2. Στην δεύτερη περίπτωση, προτείνεται ένας τρόπος αποφυγής της χρήσης του color upsampler, ρυθμίζοντας το βασικό δίκτυο να λειτουργεί στη μέγιστη ραδιομετρική ικανότητα των δεδομένων (Εξίσωση 4.2). Όπως και προηγουμένως παράγεται η απεικόνιση $I_T^{s\downarrow}$, η οποία τροφοδοτείται κατευθείαν στον core και προκύπτουν $2^8 = 256$ εντάσεις χρωμάτων




**Σχήμα 4.2:** Δεύτερο στάδιο μελέτης και τελική αρχιτεκτονική του μοντέλου CloudTran, με **δύο** δίκτυα. (καλύτερα ορατή σε μεγέθυνση)

ανά κανάλι (8-bits).

$$\begin{aligned} p(I_T|\mathcal{H}) &= p(I_T|\mathcal{H}) \cdot p(I_T^{s\downarrow}|I_T, \mathcal{H}) \\ &= p(I_T^{s\downarrow}, I_T|\mathcal{H}) \\ &= p(I_T|I_T^{s\downarrow}, \mathcal{H}) \cdot p(I_T^{s\downarrow}|\mathcal{H}^{\downarrow}) \end{aligned} \tag{4.2}$$

## 4.1 Μελέτη Αρχιτεκτονικής

Στη συνέχεια θα αναλύσουμε τη δομή κάθε δικτύου ξεχωριστά. Ξεκινώντας με τον βασικό δίκτυο ανακατασκευής, στο σενάριο που χρησιμοποιούνται και τα τρία δίκτυα, μοντελοποιεί μια κατανομή $p_c(I_T^{s\downarrow c\downarrow}|\mathcal{H}^{\downarrow})$ 512 χρωμάτων για κάθε εικονοστοιχείο, αξιοποιώντας πληροφορία από τον κύβο $\mathcal{H}^{\downarrow}$ που αναπαριστά μια χρονοσειρά καθώς και από τα ήδη παραγώμενα εικονοστοιχεία (με τη σειρά σκαναρίσματος εικόνας):

$$p_c(I_T^{s\downarrow c\downarrow}|\mathcal{H}^{\downarrow}) = \prod_{i=1}^{H^{\downarrow}} \prod_{j=1}^{W^{\downarrow}} p_c(I_T^{s\downarrow c\downarrow}(i,j)|I_T^{s\downarrow c\downarrow}(<i,\cdot), I_T^{s\downarrow c\downarrow}(i,<j), \mathcal{H}^{\downarrow}) \tag{4.3}$$

Στην περίπτωση που χρησιμοποιείται μόνο το βασικό δίκτυο, για την επίλυση του ζητήματος αναδόμησης της απεικόνισης, με χαμηλή χωρική διακριτική ικανότητα, αλλά στο πλήρες φάσμα χρωμάτων, η κατανομή που αναζητείται παίρνει τη μορφή $p_c(I_T^{s\downarrow}|\mathcal{H}^{\downarrow})$ και αντίστοιχα λαμβάνει πληροφορία όπως φαίνεται στην Εξίσωση 4.4:

$$p_c(I_T^{s\downarrow}|\mathcal{H}^{\downarrow}) = \prod_{i=1}^{H^{\downarrow}} \prod_{j=1}^{W^{\downarrow}} p_c(I_T^{s\downarrow}(i,j)|I_T^{s\downarrow}(<i,\cdot), I_T^{s\downarrow}(i,<j), \mathcal{H}^{\downarrow}) \tag{4.4}$$

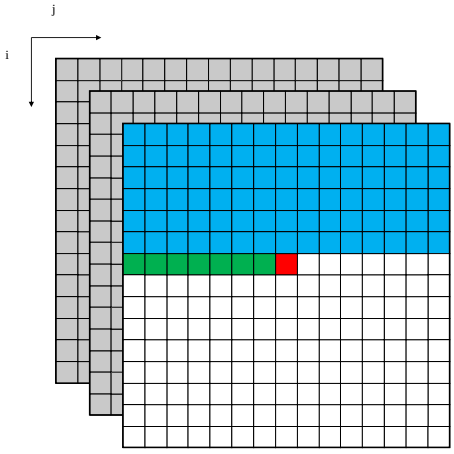
Αμφότερα και τα δύο σενάρια διαθέτουν την ίδια αρχιτεκτονική. Η αρχή γίνεται με τον Encoder, ο οποίος απαρτίζεται από 4 επίπεδα Axial Attention γραμμής και στήλης και επεξεργάζεται τη χρονοσειρά υπερφασματικών δεδομένων (κύβος $\mathcal{H}^{\downarrow}$), με μασκαρισμένες τις περιοχές επιρροής των συννέφων. Ο κύβος περνά από κατάλληλα embeddings τα οποία συμπεριλαμβάνουν και μια κωδικοποίηση θέσης (positional encoding), ώστε να ληφθεί υπόψιν η χρονική σειρά των δεδομένων. Στη συνέχεια ο Encoder παράγει πληροφορίες ${}^b\mathbf{c}_t \in \mathbb{R}^{H^{\downarrow}\times W^{\downarrow}\times B\cdot T\times D}$ για κάθε χρονική στιγμή $t$ ξεχωριστά, οι οποίες συνενώνονται με $1\times 1$ συνελικτικές στρώσεις[1], οπότε και προκύπτει μια πληροφορία ${}^b\bar{\mathbf{c}} \in \mathbb{R}^{H^{\downarrow}\times W^{\downarrow}\times D}$ για κάθε κανάλι. Τα R, G, B κανάλια σε κάθε χρονοσειρά επεξεργάζονται ανεξάρτητα μεταξύ τους. Η εξαγώμενη πληροφορία ${}^b\bar{\mathbf{c}}$ του Encoder, οριοθετεί τα επίπεδα και λειτουργεί ως επιπλέον είσοδος στα embeddings του Outer Decoder. Επιπλέον, το άθροισμα της εξόδου του Outer Decoder με τη ${}^b\bar{\mathbf{c}}$ οριοθετούν τα επίπεδα του Inner Decoder. Η δομή των Decoders είναι ίδια με αυτή του Axial Transformer, όπως περιγράφηκε στην Ενότητα 2.3. Έχουν ως είσοδο την, απαλλαγμένη από σύννεφα εικόνα $I_T^{s\downarrow}$ (ή $I_T^{s\downarrow c\downarrow}$), η οποία ταυτόχρονα αποτελεί και στόχο παραγωγής του βασικού δικτύου. Ταυτόχρονα, εκπαιδεύεται ένα ενδιάμεσο παράλληλο μοντέλο (Εξίσωση 4.5) βασισμένο αποκλειστικά στις αναπαραστάσεις που μαθαίνει ο Encoder, για λόγους ομαλοποίησης.

$$\tilde{p}_c(I_T^{s\downarrow c\downarrow}|\mathcal{H}^{\downarrow}) = \prod_{i=1}^{H^{\downarrow}} \prod_{j=1}^{W^{\downarrow}} \tilde{p}_c(I_T^{s\downarrow c\downarrow}(i,j)|\mathcal{H}^{\downarrow}) \qquad \tilde{p}_c(I_T^{s\downarrow}|\mathcal{H}^{\downarrow}) = \prod_{i=1}^{H^{\downarrow}} \prod_{j=1}^{W^{\downarrow}} \tilde{p}_c(I_T^{s\downarrow}(i,j)|\mathcal{H}^{\downarrow}) \tag{4.5}$$

Όσον αφορά την παραγωγή των αποτελεσμάτων, χρησιμοποιείται μια ημι-παράλληλη αυτο-παλινδρομητική δειγματοληψία με τη χρήση της οποίας αποφεύγεται η αξιολόγηση ολόκληρου του δικτύου για κάθε εξεταζόμενο εικονοστοιχείο (Σχήμα 4.3). Ο Encoder τρέχει μια φορά ανά κανάλι, ο Outer Decoder μια φορά ανα γραμμή της παραγώμενης απεικόνισης και ο Inner Decoder μια φορά ανα εικονοστοιχείο. Αναλυτικότερα, η πληροφορία που προέρχεται απο τον Encoder και τον Outer Decoder κατευθύνει τον Inner Decoder ώστε να παραχθεί μια γραμμή της νέας εικόνας, ανά εικονοστοιχείο. Στη συνέχεια, ο Outer Decoder επανυπολογίζει πληροφορία ώστε να τροφοδοτήσει εκ νέου τον Inner Decoder και να παραχθεί η επόμενη γραμμή.

Με σκοπό την παραγωγή των τελικών, υψηλής χωρικής διακριτικής ικανότητας, αποτελεσμάτων εκπαιδεύονται δύο ανεξάρτητα μοντέλα, ο color και ο spatial upsampler. Η αρχιτεκτονική τους είναι πανομοιότυπη, αλλά διαφέρουν στις εισόδους με τις οποίες τροφοδοτούνται και στην χωρική διακριτική ικανότητα στην οποία λειτουργούν. Παρόμοια με τον Encoder, δομούνται

[1] πειραματικά δοκιμάστηκαν και άλλες μέθοδοι συνένωσης (βλ. Υποενότητα 5.2.2)




**Σχήμα 4.3: Γκρι:** Attention γραμμης + Attention στήλης (Encoder), **Γαλάζιο:** Attention γραμμης + Μασκαρισμένο Attention στήλης (Outer Decoder), **Πράσινο:** Μασκαρισμένο Attention γραμμης (Inner Decoder), **Κόκκινο:** Παραγώμενο εικονοστοιχείο.

από 3 επίπεδα Axial Attention γραμμής και στήλης. Ο color upsampler χρησιμοποιείται μόνο στην περίπτωση του Σχήματος 4.1, όπου το βασικό δίκτυο λειτουργεί σε βάθος χρώματος 3-bits. Μετατρέπει, αρχικά, την έξοδο του βασικού δικτύου $I_T^{s\downarrow c\downarrow} \in \mathbb{R}^{H^\downarrow \times W^\downarrow \times 1}$ των 512 χρωμάτων σε μια RGB απεικόνιση με 8 εντάσεις ανά κανάλι. Στα κανάλια αυτά, εφαρμόζονται τα κατάλληλα embeddings οπότε προκύπτει ο ενδιάμεσος πίνακας $I_T^{s\downarrow c\downarrow}(b) \in \mathbb{R}^{H^\downarrow \times W^\downarrow \times D}$, όπου $b \in \{R, G, B\}$, ο οποίος αθροίζεται με τα embeddings που προκύπτουν από τα αντίστοιχα κανάλια του κύβου $\mathcal{H}^\downarrow$ και ύστερα τροφοδοτούνται στα επίπεδα του Axial Attention. Η έξοδος των στρώσεων του attention, προβάλλονται σε πιθανολογικές κατανομές, ανά κανάλι και εικονοστοχείο, $\tilde{p}_{c\uparrow}(I_T^{s\downarrow}|I_T^{s\downarrow c\downarrow}, \mathcal{H}^\downarrow) \in \mathbb{R}^{H^\downarrow \times W^\downarrow \times 256}$ για 256 εντάσεις χρωμάτων σε κάθε κανάλι. (Εξίσωση 4.6)

$$\tilde{p}_{c\uparrow}(I_T^{s\downarrow}|\mathcal{H}^\downarrow) = \prod_{i=1}^{H^\downarrow} \prod_{j=1}^{W^\downarrow} \tilde{p}_{c\uparrow}(I_T^{s\downarrow}(i,j)|I_T^{s\downarrow c\downarrow}, \mathcal{H}^\downarrow) \tag{4.6}$$

Πρώτο βήμα του spatial upsampler είναι η εφαρμογή μιας απλής χωρικής παρεμβολής στην απεικόνιση $I_T^{s\downarrow} \in \mathbb{R}^{H^\downarrow \times W^\downarrow \times 3}$. Στη συνέχεια, όπως και στον color upsampler, κάθε κανάλι μαζί με την αντίστοιχη πληροφορία από τον κύβο $\mathcal{H}$ εισάγεται στα επίπεδα του attention και ύστερα προβάλλεται ως πιθανολογική κατανομή $\tilde{p}_{s\uparrow}(I_T(b)|I_T^{s\downarrow}, \mathcal{H}) \in \mathbb{R}^{H \times W \times 256}$ για 256 εντάσεις χρωμάτων σε κάθε κανάλι. (Εξίσωση 4.7)

$$\tilde{p}_{s\uparrow}(I_T|\mathcal{H}) = \prod_{i=1}^{H} \prod_{j=1}^{W} \tilde{p}_{s\uparrow}(I_T(i,j)|I_T^{s\downarrow}, \mathcal{H}) \tag{4.7}$$

Το μοντέλο εκπαιδεύεται με σκοπό την ελαχιστοποίηση της αρνητικής λογαριθμικής πιθανοφάνειας των δεδομένων (Εξίσωση 4.8), θεωρώντας ως στόχο, την καθαρή εικόνα $I_T$. Οι όροι

$p_c/\tilde{p}_c$, $\tilde{p}_{s\uparrow}$ και $\tilde{p}_{c\uparrow}$ μεγιστοποιούνται ανεξάρτητα μεταξύ τους, ενώ το $\lambda$ είναι η υπερπαράμετρος που ελέγχει την σχετική συνεισφορά των $p_c$ και $\tilde{p}_c$.

$$\mathcal{L} = (1 - \lambda) \log p_c + \lambda \log \tilde{p}_c + \log \tilde{p}_{s\uparrow} + (\log \tilde{p}_{c\uparrow}) \tag{4.8}$$

# ΚΕΦΑΛΑΙΟ 5

# Πειράματα και Μετρικές

## 5.1 Δεδομένα

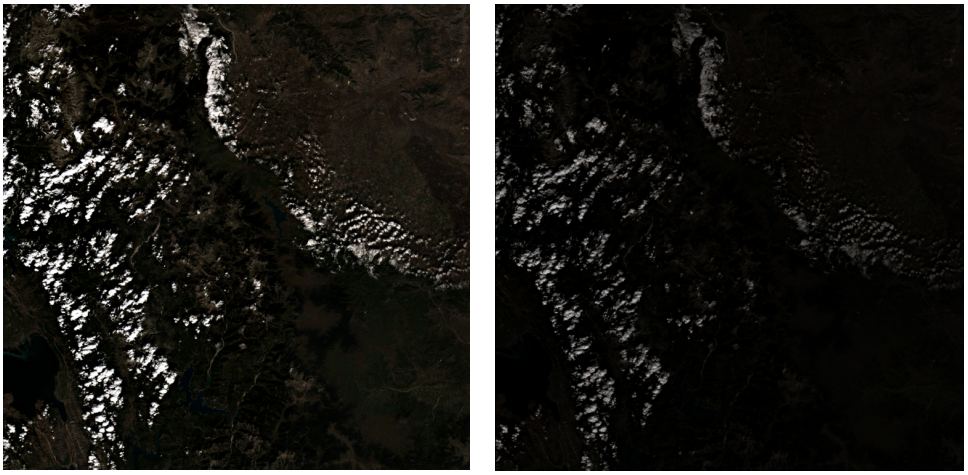
### 5.1.1 Ιδιόκτητο Σετ Δεδομένων Sentinel-2

Το σετ δεδομένων, βάση του οποίου πραγματοποιούνται τα πειράματα του αλγορίθμου, αποτελείται από 56 προϊόντα των δορυφόρων Sentinel-2 (2A/2B) επιπέδου 2A, αφορούν την ίδια περιοχή και έχουν ληφθεί την περίοδο 2018-2019. Στην παρούσα εργασία θα ασχοληθούμε με τα φασματικά κανάλια B02, B03 και B04 με χωρική διακριτική ικανότητα 10 μέτρων (Πίνακας 1.2). Από την σύνθεση των προαναφερθέντων καναλιών προκύπτει μια πολυφασματική εικόνα RGB για κάθεμια από τις 56 χρονικές στιγμές, με συνολικό μέγεθος 10.980 $\times$ 10.980 εικονοστοιχεία. Με τη σειρά της, κάθε RGB απεικόνιση διαχωρίζεται σε μικρότερες υπο-περιοχές, μεγέθους 512 $\times$ 512 εικονοστοιχείων, διαμοφώνοντας συνολικά **441 χρονοσειρές** πολυφασματικών δεδομένων. Όσον αφορά τις μάσκες συννέφων χρησιμοποιείται το κανάλι CLD του αντίστοιχου προϊόντος L2A, χωρικής διακριτικής ικανότητας 10m, το οποίο αναπαριστά την πιθανότητα ύπαρξης συννέφων για κάθε εικονοστοιχείο, με τις μη μηδενικές να απορρίπτονται και να μασκάρονται όπως θα δούμε στην επόμενη Ενότητα.

Για ευκολότερη φωτοερμηνεία των δορυφορικών απεικονίσεων, προτού πραγματοποιηθεί η σύνθεση των RGB εικόνων, οι τιμές κάθε καναλιού διαιρούνται με έναν συντελεστή, ο οποίος στην περίπτωση μας ειναι 7.000. Σκοπός είναι κάθε εικόνα $I_t$, να λαμβάνει τιμές έντασης [0,255]. Υπό κανονικές συνθήκες, η διαίρεση θα έπρεπε να γίνει με την μέγιστη τιμή των δεδομένων ($\approx$ 21.000), ωστόσο τα μεγέθη αυτά ανταποκρίνονται αποκλειστικά σε περιοχές συννέφων, οπότε πειραματικά επιλέχθηκαν οι 7.000 ως καταλληλος συντελεστής. Η οπτική σύγκριση των δύο τιμών παρουσιάζεται στο Σχήμα 5.1.

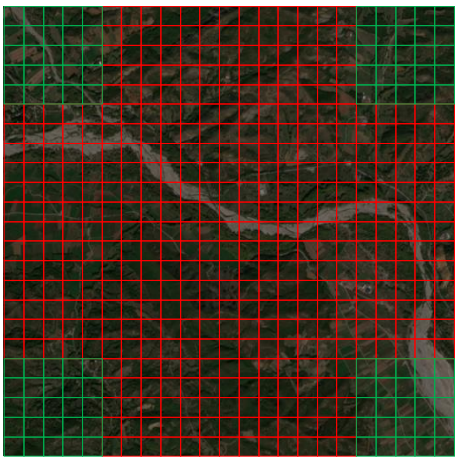
Τα δεδομένα ελέγχου, επιλέχθηκαν με ένα παράθυρο 5 $\times$ 5 των πολυφασματικών απεικονίσεων, σε κάθε μια από τις τέσσερις γωνίες τους. Συνολικά λοιπόν, προκύπτουν 5 $\times$ 5 $\times$ 4 = **100** χρονοσειρές διαφορετικών περιοχών ενδιαφέροντος, με σκοπό τον ποιοτικό έλεγχο του αλγο-

ρίθμου ($\sim$ 20% των αρχικών δεδομένων). Οι υπολοιπόμενες 341 χρονοσειρές, απαρτίζουν το σύνολο εκπαίδευσης ($\sim$ 80% των αρχικών δεδομένων). Ο διαχωρισμός συνόλου ελέγχου και εκπαίδευσης περιγράφεται οπτικά από το Σχήμα 5.2.



**(a)** **(b)**

**Σχήμα 5.1:** 5.1a: Διαίρεση αρχικών δεδομένων με συντελεστή 7.000, 5.1b: Διαίρεση αρχικών δεδομένων με συντελεστή 21.000.

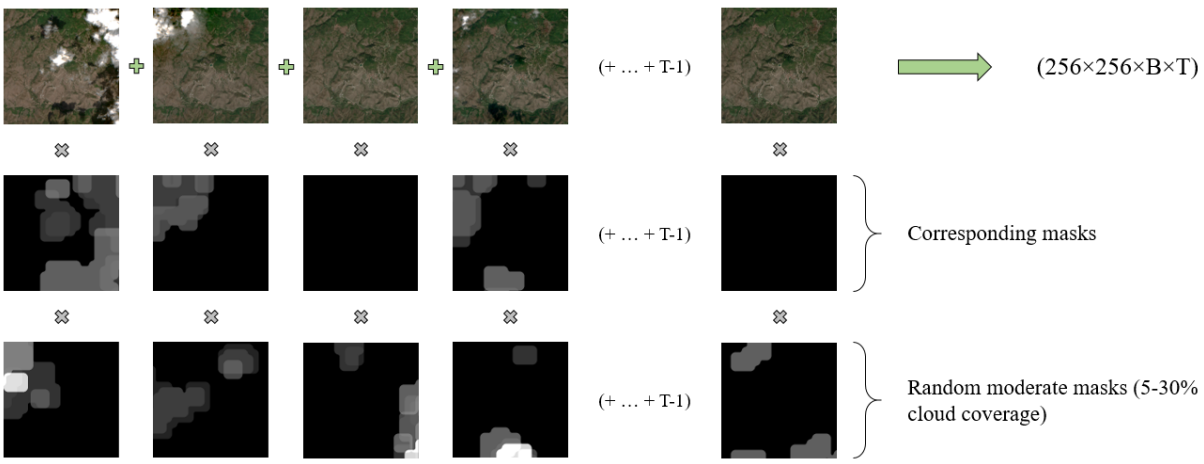


**Σχήμα 5.2:** Επιλογή δεδομένων ελέγχου (πράσινες περιοχές) και δεδομένων εκπαίδευσης (κόκκινες περιοχές).

Ως τώρα το σετ δεδομένων μας, αποτελείται από 441 χρονοσειρές, 56 χρονικών στιγμών και ανάλυση εικόνας 512 $\times$ 512 εικονοστοιχεία. Για να τροφοδοτηθούν όμως, τα δίκτυα του μοντέλου, τα δεδομένα χρειάζονται περαιτέρω προετοιμασία η οποία πραγματοποιείται στο εσωτερικό του αλγορίθμου.

Δημιουργείται ο κύβος $\mathcal{H} \in \mathbb{R}^{H \times W \times B \times T}$ (Σχήμα 5.3) όπου $B$ ο αριθμός των πολυφασματικών καναλιών κάθε εικόνας και $T = 5$ χρονικές στιγμές στην περίπτωση μας, ωστόσο ο συντελε-

στής αυτός μπορεί να μεταβληθεί, βάση της διατιθέμενης υπολογιστικής ικάνότητας. Τα $H, W$ που αντιπροσωπεύουν το ύψος και το πλάτος των δεδομένων αντίστοιχα, ορίζονται 256×256 προκύπτωντας από τις αρχικές 512×512 εικόνες με τυχαία περικοπή για την εκπαίδευση και κεντρική περικοπή για τον έλεγχο του μοντέλου. Για όλες τις εικόνες του κύβου εφαρμόζεται η μάσκα συννέφων που αντιστοιχεί σε κάθε χρονική στιγμή. Επιπλέον, αποθηκεύεται μια λίστα με μάσκες μέσης κάλυψης συννέφων (5-30%), μερικές από τις οποίες εφαρμόζονται τυχαία σε κάθε εικόνα του κύβου $\mathcal{H}$ κατά την διάρκεια της εκπαίδευσης. Ο έλεγχος του μοντέλου γίνεται βάση της τελευταίας χρονικά εικόνας $I_T^{GT}$ του κύβου, η οποία πρέπει να είναι απαλλαγμένη από την επιρροή συννέφων (<5%) πριν την εφαρμογή της τυχαίας μάσκας. Σημειώνεται πως ο κύβος $\mathcal{H}$ για να εισαχθεί στον encoder του βασικού δικτύου ανακατασκευής και του color upsampler (στην περίπτωση που αυτός χρησιμοποιηθεί) περνά από μια τελευταία διεργασία κατά την οποία όλες οι απεικονίσεις οδηγούνται σε χαμηλότερη διακριτική ικανότητα σχηματίζοντας τον $\mathcal{H}^{\downarrow} \in \mathbb{R}^{64\times64\times B\times T}$.




**Σχήμα 5.3:** Δημιουργία κύβου $\mathcal{H} \in \mathbb{R}^{H\times W\times B\times T}$

### 5.1.2 Δημόσιο Σετ Δεδομένων SEN12MS-CR-TS

Το SEN12MS-CR-TS δημοσιεύτηκε το 2022 [19] και συντίθεται από χρονοσειρές εικόνων SAR και πολυφασματικών δεδομένων τα οποία καταγράφονται από τους δορυφόρους Sentinel-1 και Sentinel-2 αντίστοιχα, σε παγκόσμια κλίμακα. Για τους σκοπούς της παρούσας εργασίας χρησιμοποιούνται μόνο τα πολυφασματικά δεδομένα Sentinel-2 (επιπέδου 1C) και πιο συγκεκριμένα τα κανάλια B02, B03 και B04 με χωρική διακριτική ικανότητα 10 μέτρων για την παραγωγή των RGB έγχρωμων συνθέτων. Παρέχονται συνολικά 43 περιοχές ενδιαφέροντος, κάθε μια από τις οποίες μελετάται για 30 χρονικές στιγμές ομοιόμορφα κατανεμημένες κατά τη διάρκεια του έτους 2018, ενώ το χρονικό διάστημα των διαδοχικών παρατηρήσεων δεν ξεπερνά τις δύο εβδομάδες. Κάθε περιοχή ενδιαφέροντος καλύπτει περίπου $40 \times 40\ km^2$ εδάφους ($\approx 4000 \times 4000\ px^2$)

και χωρίζεται σε μη επικαλυπτόμενα τμήματα διαστάσεων $256 \times 256\ px^2$ με αποτέλεσμα ο τελικός αριθμός των διαθέσιμων περιοχών να ανέρχεται στις $12293$. Ο διαχωρισμός σε δεδομένα εκπαίδευσης και ελέγχου γίνεται βάση του [19] με αποτέλεσμα $11262$ περιοχές να αποτελούν το σετ εκπαίδευσης και $1031$ το σετ ελέγχου/επαλήθευσης.

Με σκοπό την παραγωγή δυαδικών μασκών, οι οποίες δεν παρέχονται στο παρών σετ δεδομένων, χρησιμοποιείται ο αλγόριθμος s2cloudless που είναι διαθέσιμος μέσω του Sentinel-Hub. Δέχεται για εισόδους τις Sentinel-2 εικόνες, διαστάσεων H×W×B όπου B τα φασματικα κανάλια του προϊόντος. Σε αυτό το σημείο υπάρχει η επιλογή χρήσης είτε και των 13 καναλιών κάθε εικόνας ή 10 εξ αυτών χωρίς τα B03, B06 και B07. Ο αλγόριθμος λειτουργεί σε επίπεδο εικονοστοιχείου υπολογίζοντας την πιθανότητα κατηγοριοποίησης του ως σύννεφο. Η έξοδος του είναι μια δυαδική μάσκα για κάθε εικόνα, με τα εικονοστοιχεία που αναγνωρίζονται ως σύννεφα να λαμβάνουν την τιμή 1, ενώ τα καθαρά εικονοστοιχεία την τιμή 0. ΄Υστερα από πειράματα, η πιθανότητα 60% επιλέγεται ως κατώφλι πάνω από το οποίο το εικονοστοιχείο κατηγοριοποιείται ως σύννεφο. Σημειώνεται πως, τόσο για την παραγωγή των μασκών όσο και για την είσοδο τους στα δίκτυα του CloudTran, οι αρχικές τιμές ανακλαστικότητας των δεδομένων (16-bits) διαιρούνται με τον προτεινόμενο συντελεστή 10000 ώστε οι τελικές τιμές να εντοπίζονται στο εύρος [0,255] (8-bits).

Για την δημιουργία του κύβου $\mathcal{H} \in \mathbb{R}^{H \times W \times B \times T}$ ακολουθείται η ίδια διαδικασία με αυτή που περιγράφηκε στην Υποενότητα 5.1.1. Η μόνη διαφορά έγκειται στις διαστάσεις των επιμέρους τμημάτων κάθε περιοχής ενδιαφέροντος, οι οποίες για το σετ δεδομένων SEN12MS-CR-TS είναι 256×256. Η αρχική περικοπή, λοιπόν, οδηγεί σε διαστάσεις $224 \times 224\ px^2$ και ύστερα ελαττώνεται η διακριτική ικανότητα κάθε απεικόνισης σε $64 \times 64\ px^2$.

## 5.2 Πειράματα

Στην Ενότητα αυτή θα περιγραφούν όλες οι δομές πειραμάτων που εφαρμόστηκαν, σε χρονική σειρά, μέχρι να καταλήξουμε στην τελική αρχιτεκτονική (Σχήμα 4.2) καθώς και οι διαφορές στις παραμέτρους, τα πλεονεκτήματα και τα μειονεκτήματά τους. Τέλος, πραγματοποιείται μελέτη της σημασίας ορισμένων σημαντικών υπερπαραμέτρων, για την ομαλή λειτουργία του αλγορίθμου, η οποία θα παρουσιαστεί μέσω πινάκων και γραφημάτων.

### 5.2.1 Σταθερές παράμετροι μοντέλου

Όλα τα τελικά πειράματα εκπαίδευσης του βασικού δικτύου ανακατασκευής, του color upsampler και του spatial upsampler πραγματοποιήθηκαν ανεξάρτητα, σε σύστημα με δύο κάρτες γραφικών της NVIDIA ***Quadro RTX 6000***, με διαθέσιμη μνήμη VRAM ***24GB*** η καθεμία.

| **Core Parameters** | |
|---|---|
| Encoder layers | $4$ |
| Outer Decoder layers | $2$ |
| Inner Decoder layers | $2$ |
| Attention heads | $4$ |
| Batch size | $1\ (hypercube\ \mathcal{H})$ |
| Max train steps | $15000$ |
| Optimizer | $RMSProp$ |
| Learning rate | $3e^{-4}$ |
| Decay value | $0.999$ |

**Πίνακας 5.1:** Υπερπαράμετροι βασικού δικτύου ανακατασκευής (Core).

Στους Πίνακες 5.1 - 5.3 παρουσιάζονται οι υπερπαράμετροι του βασικού δικτύου, του color upsampler και του spatial upsampler αντίστοιχα, οι οποίοι κατά την διάρκεια όλων των πειραμάτων παραμένουν σταθεροι εκτός αν, σε μεμονομένα πειράματα, αναφέρεται διαφορετικά. Κάθε μοντέλο του βασικού δικτύου και του spatial upsampler εκπαιδεύεται για $15000$ επαναλήψεις, χρησιμοποιώντας τον αλγόριθμο βελτιστοποίησης RMSProp με συντελεστή μάθησης $3 \cdot 10^{-4}$. Το ίδιο ισχύει και για τα μοντέλα του color upsampler με τη μόνη διαφοροποίηση ότι εκπαιδεύονται για $30000$ επαναλήψεις. Αξίζει να σημειωθεί ότι στο βασικό δίκτυο, τα σχετικά βάρη μεταξύ των λογαριθμικών απωλειών του decoder και του encoder είναι $0.99$ και $0.01$ αντίστοιχα.

| Color Upsampler Parameters | |
|---|---|
| Encoder layers | $3$ |
| Attention heads | $4$ |
| Batch size | $1\ (hypercube\ \mathcal{H})$ |
| Max train steps | $30000$ |
| Optimizer | $RMSProp$ |
| Learning rate | $3e^{-4}$ |
| Decay value | $0.999$ |

**Πίνακας 5.2:** Υπερπαράμετροι δικτύου Color Upsampler.

| Spatial Upsampler Parameters | |
|---|---|
| Encoder layers | $3$ |
| Attention heads | $4$ |
| Batch size | $1\ (hypercube\ \mathcal{H})$ |
| Max train steps | $15000$ |
| Optimizer | $RMSProp$ |
| Learning rate | $3e^{-4}$ |
| Decay value | $0.999$ |

**Πίνακας 5.3:** Υπερπαράμετροι δικτύου Spatial Upsampler.

### 5.2.2 Περιγραφή Δομών Πειραμάτων

Στις δομές που ακολουθούν μελετάται η αρχιτεκτονική και η χρήση του βασικού δικτύου και του color upsampler καθώς αυτά αναλαμβάνουν την αναδόμηση της νέας, απαλλαγμένης από σύννεφα απεικόνισης, στις επιλεγμένες διαστάσεις υποδειγματογράφησης (π.χ 64×64). Το δίκτυο του spatial upsampler αναλάμβάνει την μετάβαση της παραγόμενης απεικόνισης των προηγούμενων δικτύων στις αρχικές διαστάσεις μετά την περικοπή (256×256 για το ιδιόκτητο σετ δεδομένων ή 224×224 για το SEN12MS-CR-TS), οπότε η χρήση του είναι απαραίτητη σε κάθε δοκιμή με σκοπό την παραγωγή των τελικών αποτελεσμάτων.

### Δομή 1

Ξεκινώντας τα πειράματα, προσαρμόσαμε το μοντέλο του Colorization Transformer[1] [17] ώστε να είναι συμβατό με τις χρονοσειρές RGB απεικονίσεων που διαθέτουμε. Είσοδος του Encoder, αποτελεί ο κύβος $\mathcal{H} \in \mathbb{R}^{64\times64\times3\times5}$, με **3 κανάλια** για κάθε απεικόνιση (B02, B03, B04) και **5 συνολικά απεικονίσεις** διαφορετικών χρονικών στιγμών. Όπως περιγράφηκε στην Ενότητα 5.1, οι τελικές διαστάσεις 64×64 της εισόδου προκύπτουν με τυχαία ή κεντρική περικοπή των αρχικών δεδομένων, από 512×512 σε 256×256 και με υποδειγματογράφηση στο τελικό μέγεθος. Είσοδος του Decoder αποτελεί η πέμπτη (τελευταία χρονικά) εικόνα, η οποία είναι απαλλαγμένη από την επιρροή συννέφων, με βάθος χρώματος από 8 σε **3 bits**. Στη συνέχεια χρησιμοποιείται ο color upsampler, με σκοπό την αύξηση του βάθους χρωμάτων της νέας αναδομημένης RGB εικόνας, που προκύπτει από το βασικό δίκτυο, σε 8-bits με 256 εντάσεις ανά κανάλι.

*Με αυτή τη δομή καλούμαστε να εκπαιδεύσουμε **2 διαφορετικά μοντέλα**, έναν core και έναν color upsampler, ανεξάρτητα μεταξύ τους.*

### Δομή 2

Δημιουργούνται τρία νέα σετ δεδομένων, με τη μοναδική διαφορά ότι πλέον κάθε ένα περιέχει απεικονίσεις των καναλιών B02, B03, B04 του Sentinel-2 ξεχωριστά. Αυτό σημαίνει πως η είσοδος στον Encoder, είναι ένας κύβος $\mathcal{H}_b \in \mathbb{R}^{64\times64\times1\times5}$, όπου $b \in \{R, G, B\}$, με συνολικά **5 απεικονίσεις** της ίδιας περιοχής σε διαφορετικές χρονικές στιγμές, για **1 κανάλι** κάθε φορά. Όλα τα μοντέλα εκπαιδεύονται ανεξάρτητα για κάθε κανάλι με την ίδια ακριβώς αρχιτεκτονική της Δομής 1.

*Η προσέγγιση αυτή, δεν είναι χρήσιμη, καθώς καλούμαστε να εκπαιδεύσουμε **6 διαφορετικά μοντέλα**, τρεις πυρήνες και τρεις color upsamplers, για την παραγωγή μιας αναδομημένης RGB εικόνας.*

### Δομή 3

Όπως και προηγουμένως, χρησιμοποιούνται τα σετ δεδομένων ενός καναλιού, οπότε η είσοδος του Encoder είναι ο κύβος $\mathcal{H}_b$ που είδαμε και στην Δομή 2. Η σημαντική διαφοροποίηση σε αυτή την περίπτωση, έγκειται στα δεδομένα εισόδου και τη λειτουργία του Decoder. Στα προηγούμενα πειράματα είδαμε πως η ground truth εικόνα ($I_T^{GT}$) προκειμένου να εισαχθεί στον Decoder, υπόκειται μείωση στο βάθος του χρώματος της από 8-bits σε 3-bits και η τελική έξοδος του βασικού δικτύου είναι μια πιθανολογική κατανομή 8 εντάσεων ανά κανάλι. Με την παρούσα προσέγγιση η απεικόνιση $I_T^{GT}$ διατηρεί το αρχικό βάθος χρώματος των **8-bits** και ο Decoder λειτουργεί με σκοπό την εύρεση μιας πιθανολογικής κατανομής για 256 εντάσεις ανά

[1] https://github.com/google-research/google-research/tree/master/coltran

κανάλι. Βρισκόμαστε λοιπόν στην περίπτωση του Σχήματος 4.2 όπου δεν χρησιμοποιείται ο color upsampler, αλλά μόνο το βασικό δίκτυο ανακατασκευής. Όλα τα μοντέλα εκπαιδεύονται ανεξάρτητα για κάθε κανάλι και ύστερα, τα αποτελέσματά τους συνδυάζονται για την δημιουργία της καθαρής αναδομημένης RGB εικόνας.

*Με αυτή τη δομή καλούμαστε να εκπαιδεύσουμε **3 διαφορετικά μοντέλα**, τρεις πυρήνες συγκεκριμένα.*

Ο αριθμός των απαιτούμενων μοντέλων προς εκπαίδευση, με αυτή τη προσέγγιση, είναι μεγαλύτερος από αυτόν της Δομής 1 (3 και 2 αντίστοιχα), αλλά έχει το πλεονέκτημα πως κάθε μοντέλο καταναλώνει λιγότερους υπολογιστικούς πόρους. Στα μειονεκτήματα της μεθόδου, κάθε κανάλι αναδομείται ανεξάρτητα από τα υπόλοιπα δύο, οπότε είναι πιθανή η ύπαρξη χοντρικών σφαλμάτων στις τιμές ορισμένων εικονοστοιχείων της τελικής απεικόνισης.

**ΧΡΗΣΗ ΦΙΛΤΡΟΥ ΕΝΔΙΑΜΕΣΗΣ ΤΙΜΗΣ**

Στο σημείο αυτό, έγινε ένα ενδιάμεσο πείραμα με την ίδια αρχιτεκτονική αλλά με ένα νέο σετ δεδομένων, το οποίο περιέχει χρονοσειρές απεικονίσεων ήδη απαλλαγμένων από σύννεφα, με τη χρήση του φίλτρου ενδιάμεσης τιμής. Για την προετοιμασία των δεδομένων, δημιουργήθηκε ένα κινούμενο παράθυρο αποτελούμενο από 5 απεικονίσεις στον άξονα του χρόνου κατά το οποίο εφαρμόζεται το φίλτρο median. Είσοδος του Encoder είναι ο κύβος $\mathcal{H}_b \in \mathbb{R}^{64\times64\times1\times7}$, που περιέχει απεικονίσεις του εκάστοτε καναλιού $b \in \{R, G, B\}$ της ίδιας περιοχής ενδιαφέροντος για 7 χρονικές στιγμές. Σε κάθε εικόνα του $\mathcal{H}_b$ αντιστοιχίζεται μια τυχαία μάσκα με ποσοστό νεφοκάλυψης 5-30%. Το πείραμα αυτό έχει σκοπό να δοκιμάσει την δύναμη και το βάθος του αλγορίθμου, καθώς οι περιοχές επιρροής των συννέφων αναδομούνται πρόχειρα με το φίλτρο ενδιάμεσης τιμής και στη συνέχεια δεν μασκάρονται, οπότε το μοντέλο καλείται να «μάθει» την πραγματική κατανομή που κρύβεται πίσω από αυτές.

### Δομή 4 (Τελική)

Σε εξέλιξη των προηγούμενων πειραμάτων, προσαρμόζεται το βασικό δίκτυο με σκοπό να επεξεργάζεται **ταυτόχρονα και ανεξάρτητα** τα τρία κανάλια κάθε απεικόνισης με τη χρήση κατάλληλων positional encodings. Αυτό σημαίνει πως χρησιμοποιείται το αρχικό σετ δεδομένων με τις RGB εικόνες και κατ' επέκταση είσοδο του Encoder αποτελεί ο κύβος $\mathcal{H} \in \mathbb{R}^{64\times64\times3\times5}$. Όσον αφορά τον Decoder, λειτουργεί στο πλήρες φάσμα χρωμάτων (8-bits) των πολυφασματικών δεδομένων οπότε δεν γίνεται χρήση του color upsampler (Σχήμα 4.2).

*Με αυτή τη δομή καλούμαστε να εκπαιδεύσουμε μόνο **1 μοντέλο** και συγκεκριμένα ένα δίκτυο core. Από αυτό το σημείο και ύστερα η μελέτη των υπερπαραμέτρων θα έχει ως αναφορά αυτή τη δομή και ως μεγέθη σύγκρισης, τα αποτελέσματα της συνάρτησης απώλειας κατά τον έλεγχο*

*του εκάστοτε μοντέλου (validation losses).*

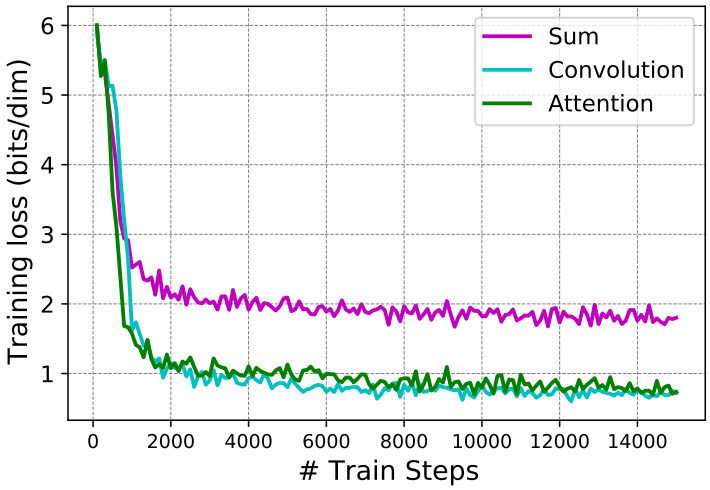
Αρχικά, μελετάται ο τρόπος συνένωσης της πληροφορίας ${}^b\mathbf{c}_t$ που παράγει ο Encoder για κάθε χρονική στιγμή και κανάλι του κύβου $\mathcal{H}$. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.4 για τον έλεγχο και μέσω γραφήματος στο Σχήμα 5.4 για την εκπαίδευση του μοντέλου, όπου η μέθοδος *Sum* αναφέρεται στην απλή πρόσθεση της πληροφορίας κάθε καναλιού στον άξονα του χρόνου, η μέθοδος *Attention* στην εφαρμογή ενός επιπέδου self-attention επίσης στον άξονα του χρόνου και τέλος η μέθοδος *Convolution* που αντιστοιχεί στην εφαρμογή $1\times1$ συνέλιξης στην εξαγόμενη πληροφρορία ${}^b\mathbf{c}_t$. Είναι εμφανές, πως η σύμπτυξη του άξονα του χρόνου μέσω απλής πρόσθεσης των πληροφοριών που λαμβάνει το μοντέλο ανά χρονική στιγμή, που είναι η συνήθης μέθοδος παραμετροποίησης των επιπέδων του Decoder, δεν είναι αρκετή ώστε να τροφοδοτήσει επαρκώς το υπόλοιπο δίκτυο με χρήσιμη πληροφορία. Αντιθέτως η χρήση επιπέδων μάθησης όπως η συνέλιξη και το self-attention οδηγούν σε σαφώς βελτιωμένα αποτελέσματα κατά μια τάξη μεγέθους. Για την σύγκριση των δύο τελευταίων μεθόδων, οι τιμές ελαχιστοποίησης της συνάρτησης απώλειας είναι πολύ κοντά μεταξύ τους χωρίς να επιφέρουν ποιοτικές αλλαγές στις εξόδους του μοντέλου. Ωστόσο προτιμότερη κρίνεται η χρήση της **$1\times1$ συνέλιξης** καθώς εισάγει σημαντικά λιγότερες παραμέτρους από τα επίπεδα attention.

| **Context Aggregation** | **Validation Loss** |
|---|---|
| Sum | 1.273 |
| Attention | 0.239 |
| **Convolution** | **0.206** |

**Πίνακας 5.4:** Σύγκριση μεθόδων συνένωσης της εξαγόμενης πληροφορίας του Encoder.

Στη συνέχεια ερευνάται η επιρροή του μεγέθους του μοντέλου στα τελικά αποτελέσματα του βασικού δικτύου. Συγκεκριμένα, στον Πίνακα 5.5 παρουσιάζονται οι τελικές τιμές της συνάρτησης απώλειας, κατά τον έλεγχο του μοντέλου, για μεγέθη 64, 128, 256 και 512. Η υπερπαράμετρος αυτή αναφέρεται στο μέγεθος που χρησιμοποιείται στο εσωτερικό των επιπέδων attention και των δικτύων τροφοδοσίας, καθώς και κατά την είσοδο στα επίπεδα ενσωμάτωσης (embeddings). Παρατηρούμε, πως η αύξηση του μεγέθους από 64 σε 128 βελτιώνει την ποιότητα αναδόμησης των τελικών απεικονίσεων, ωστόσο μια περαιτέρω αύξηση σε 256 ή 512 οδηγεί σε πολύ σημαντική υποβάθμιση των αποτελεσμάτων. Το γεγονός αυτό, αποδίδεται στον πολύ μεγάλο αριθμό παραμέτρων του μοντέλου κατά την εκπαίδευση και ταυτόχρονα στη χρήση ενός σχετικά μικρού συνόλου δεδομένων.

Η ίδια σύγκριση πραγματοποιείται και για το δίκτυο του spatial upsampler στον Πίνακα 5.6. Σε αυτή την περίπτωση αυξάνοντας το μέγεθος του μοντέλου αυξάνεται σταδιακά και η ποιότητα




**Σχήμα 5.4:** Τιμές συνάρτησης απώλειας κατά την εκπαίδευση των μοντέλων με διαφορετικές μεθόδους συνένωσης της εξαγόμενης πληροφορίας του Encoder.

| Model size | # Parameters | Validation Loss |
|---|---|---|
| 64 | 0.9M | 0.224 |
| **128** | **3.2M** | **0.206** |
| 256 | 12.0M | 3.44 |
| 512 | 46.4M | 5.54 |

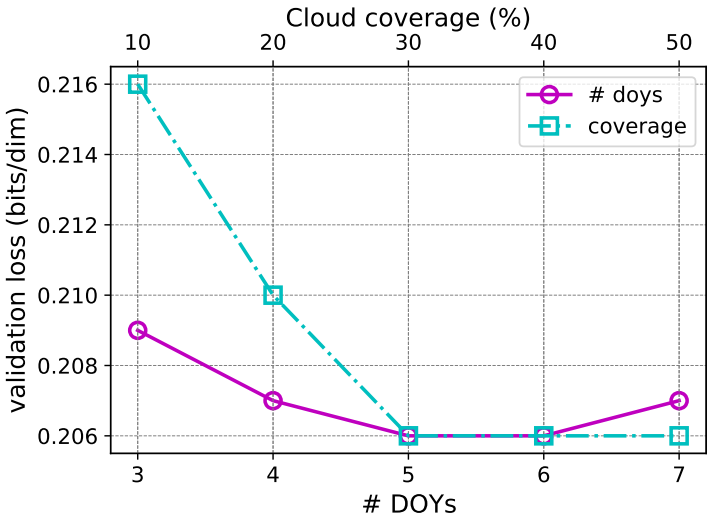
**Πίνακας 5.5:** Σύγκριση διαφορετικών μεγεθών μοντέλου για το βασικό δίκτυο ανακατασκευής.

των ενισχυμένων τελικών απεικονίσεων. Με δεδομένο πως το δίκτυο του spatial upsampler λειτουργεί με εισόδους μεγαλύτερων διαστάσεων και πως η υπολογιστική πολυπλοκότητα των επιπέδων axial attention είναι $\mathcal{O}(N\sqrt{N})$, η αύξηση του μεγέθους του μοντέλου αντιστοιχεί σε ανάλογη αύξηση των απαιτούμενων πόρων μνήμης VRAM. Έχοντας υπόψιν τον περιορισμό αυτό, το μεγαλύτερο μέγεθος στο οποίο μπορούμε να εκπαιδεύσουμε το μοντέλο είναι 512.

| Model size | # Parameters | Validation Loss |
|---|---|---|
| 64 | 0.5M | 0.266 |
| 128 | 1.3M | 0.251 |
| 256 | 3.9M | 0.246 |
| **512** | **12.7M** | **0.241** |

**Πίνακας 5.6:** Σύγκριση διαφορετικών μεγεθών μοντέλου για το δίκτυο του spatial upsampler.

Τέλος, γίνονται δοκιμές όσον αφορά τον αριθμό των χρονικών στιγμών $T$ που απαρτίζουν τον κύβο $\mathcal{H}$ καθώς και το μέγιστο επιτρεπόμενο ποσοστό νεφοκάλυψης των μασκών, που επιβάλλονται τυχαία σε κάθε χρονική στιγμή, κατά τη διάρκεια της εκπαίδευσης. Τα αντίστοιχα αποτε-




**Σχήμα 5.5:** Σύγκριση τιμών της συνάρτησης απώλειας κατά τον έλεγχο του μοντέλου, βάση του διαφορετικού αριθμού χρονικών στιγμών ως δεδομένα εισόδου και διαφορετικού μέγιστου επιτρεπόμενου ποσοστού νεφοκάλυψης κατά την εκπαίδευση.

λέσματα παρουσιάζονται, με μορφή γραφήματος, στο Σχήμα 5.5. Παρατηρούμε πως η αύξηση του χρονικού ορίζοντα των δεδομένων εισόδου από 3 σε 5 οδηγεί σε μείωση τιμών της συνάρτησης απώλειας, κατά τον έλεγχο του μοντέλου, οπότε και σε βελτιωμένα αποτελέσματα. Αυξάνοντας ακόμα περισσότερο τις διαθέσιμες χρονικές στιγμές, το μοντέλο ξεκινά να επηρεάζεται αρνητικά, γεγονός που μπορεί να αποδοθεί στις εποχικές αλλαγές που υφίστανται οι περιοχές ενδιαφέροντος κατά τη διάρκεια του έτους. Αναφορικά με το μέγιστο ποσοστό νεφοκάλυψης που ορίζεται για την εκπαίδευση, όσο μεγαλύτερο είναι (εως $50\%$), τόσο πιο θετικά επηρεάζει τις παραγόμενες απεικονίσεις, γεγονός αρκετά εύλογο αφού το μοντέλο γίνεται ολοένα και πιο αποτελεσματικό στην ανοικοδόμηση μεγαλύτερων περιοχών επηρεασμένων από σύννεφα. Οι διακυμάνσεις των τιμών της συνάρτησης απώλειας και στις δύο περιπτώσεις πειραμάτων, είναι πολύ μικρές, αποδεικνύοντας την αξιοπιστία της αρχιτεκτονικής του αλγορίθμου CloudTran.

## 5.3 Πειραματικά Αποτελέσματα και Μετρικές Αξιολόγησης

### 5.3.1 Θεωρητικό Υπόβαθρο

Η πρώτη και απλούστερη μετρική που χρησιμοποιείται για την εκτίμηση της ποιότητας του μοντέλου είναι το **Μέσο Τετραγωνικό Σφάλμα** (MSE - Mean Square Error) και υπολογίζεται ως η μέση τιμή του τετραγώνου των διαφορών μεταξύ των παραγόμενων, από το μοντέλο, τιμών και των πραγματικών. (Εξίσωση 5.1)

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (Y_{i,j} - \hat{Y}_{i,j})^2 \tag{5.1}$$

όπου $H \times W$ το μέγεθος της απεικόνισης, $\hat{Y}_{i,j}$ η τιμή της αναδομημένης και $Y_{i,j}$ η τιμή της ground truth εικόνας για το εικονοστοιχείο της θέσης $(i, j)$. Το πλεονέκτημα του Μέσου Τετραγωνικού Σφάλματος είναι πως εντοπίζει και δίνει μεγάλη βαρύτητα σε ακραίες τιμές προβλέψεων, με μεγάλο σφάλμα, λόγω του τετραγώνου στην Εξίσωση 5.1. Ο εντοπισμός αυτών των σημαντικών λαθών αποτελεί ζητούμενο της προσέγγισης μας, αλλά στον αντίποδα η μεγέθυνση τους οδηγεί σε μεγάλο τελικό σφάλμα και ας είναι μικρό το πλήθος τους. Η παραπάνω εξίσωση αναφέρεται σε ένα κανάλι, οπότε για την εύρεση του Μέσου Τετραγωνικού Σφάλματος μιας RGB εικόνας, υπολογίζεται ο μέσος όρος των σφαλμάτων των τριών καναλιών.

Η δεύτερη μετρική που χρησιμοποιείται ονομάζεται **PSNR** (Peak Signal-to-Noise Ratio) και είναι η πιο διαδεδομένη μέτρηση για την αξιολόγηση της ποιότητας μιας αναδομημένης εικόνας. Εκφράζει την αναλογία μεταξύ της μέγιστης δυνατής τιμής του σήματος και της ισχύος του θορύβου που επηρεάζει την απεικόνιση.

$$PSNR = 10 \log_{10} \left( \frac{(2^n - 1)^2}{MSE} \right) \tag{5.2}$$

όπου $n$ ο παράγοντας που καθορίζει το βάθος χρώματος της εικόνας (n=8 για τιμές εντάσεων [0.255]). Όσο μεγαλύτερη είναι η τιμή του PSNR, τόσο καλύτερη είναι η ποιότητα της αναδομημένης απεικόνισης και η ομοιότητα της με την ground truth διότι ελαχιστοποιείται το MSE συγκριτικά με τη μέγιστη τιμή του σήματος.

Η τρίτη και τελευταία μετρική είναι η **SSIM** (Structural Similarity Index Measure) [1], η οποία αξιολογεί την ομοιότητα μεταξύ δύο εικόνων. Η σύγκριση τους πραγματοποιείται λαμβάνοντας υπόψιν 3 βασικά χαρακτηριστικά, τη Φωτεινότητα (Luminance), την Αντίθεση (Contrast) και την Δομή (Structure) της απεικόνισης μεσω των εξισώσεων 5.3, 5.4, 5.5 αντίστοιχα.

$$l(Y,\hat{Y}) = \frac{2\mu_Y\mu_{\hat{Y}} + C_1}{\mu_Y^2 + \mu_{\hat{Y}}^2 + C_1} \tag{5.3}$$

$$c(Y,\hat{Y}) = \frac{2\sigma_Y\sigma_{\hat{Y}} + C_2}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + C_2} \tag{5.4}$$

$$s(Y,\hat{Y}) = \frac{\sigma_{Y\hat{Y}} + C_3}{\sigma_Y\sigma_{\hat{Y}} + C_3} \tag{5.5}$$

όπου $C_1, C_2, C_3$ σταθεροί αριθμοί για την αποφυγή διαίρεσης με παρανομαστή μηδέν, $\mu, \sigma$ ο μέσος όρος και η τυπική απόκλιση της εικόνας και $\sigma_{Y\hat{Y}}$ ο συντελεστής συσχέτισης μεταξύ ground truth και αναδομημένης απεικόνισης. Η τελική μορφή του SSIM περιγράφεται στην Εξίσωση 5.6 και λαμβάνει τιμές από 0 έως 1. Όσο μεγαλύτερη είναι η τιμή της μετρικής τόσο μεγαλύτερη η ομοιότητα των δύο εικόνων, ενώ αν ισούται με 1 σημαίνει ότι ταυτίζονται.

$$SSIM = l(Y,\hat{Y}) \cdot c(Y,\hat{Y}) \cdot s(Y,\hat{Y}) \tag{5.6}$$

### 5.3.2 Ποσοτικά και ποιοτικά αποτελέσματα

Λαμβάνοντας υπόψιν όλες τις εκδοχές πειραμάτων της Ενότητας 5.2, με σκοπό την παραγωγή των τελικών αποτελεσμάτων και τη σύγκριση του αλγορίθμου μας με άλλες μεθόδους, βασιζόμαστε σε ένα μοντέλο το οποίο εκπαιδεύει ένα βασικό δίκτυο ανακατασκευής μεγέθους 128, με διαστάσεις υποδειγματογράφησης 64×64, ακολουθούμενο από ένα δίκτυο spatial upsampler μεγέθους 512 και συνδυαζόμενα αναδομούν την τελική, απαλλαγμένη από σύννεφα, απεικόνιση διαστάσεων 256×256 (ή 224×224 ανάλογα με το σετ δεδομένων που μελετάται).

Ο Πίνακας 5.7 συγκρίνει την απόδοση της μεθόδου CloudTran στις ενδιάμεσες διαστάσεις 64×64 που προκύπτουν από το βασικό δίκτυο, με διάφορες άλλες μεθόδους αναδόμησης. Συγκεκριμένα, προσαρμόζεται στα δεδομένα και τις απαιτήσεις μας η μέθοδος αναδόμησης βίντεο FFVI [9] καθώς και μια πρότυπη μέθοδος γεμίσματος κενών που βασίζεται σε κινούμενο φίλτρο ενδιάμεσης τιμής. Επιπλέον, επιλέγεται για σύγκριση και η μέθοδος SpAGAN [13], ενδεικτική για περιπτώσεις αναδόμησης απεικονίσεων απαλλαγμένων από σύννεφα χωρίς την χρήση χρονοσειρών, αλλά μιας αποκλειστικά χρονικής στιγμής. Το γεγονός αυτό, ευθύνεται για την υψηλή τιμή του Μέσου Τετραγωνικού Σφάλματος (MSE) που αντιστοιχεί στη μέθοδο SpAGAN. Με την ίδια φιλοσοφία δομείται ο Πίνακας 5.8 στον οποίο παρουσιάζονται οι μετρικές της πλήρους αρχιτεκτονικής CloudTran, ύστερα δηλαδή και από την χρήση του δικτύου spatial upsampler και παραγωγή των τελικών απεικονίσεων 256×256.

Αμφότεροι οι δύο πίνακες παρουσιάζουν επιπλέον, τα αποτελέσματα της μεθόδου CloudTran που προκύπτουν αποκλειστικά από τις εξόδους του Encoder με τον χαρακτηρισμό CloudTran

Parallel. Όντας παράλληλο μοντέλο, η δειγματοληψία των τελικών αποτελεσμάτων είναι πολύ γρηγορότερη, ωστόσο η ποιότητα των παραγόμενων απεικονίσεων είναι επηρεασμένη από μεμονομένες αστοχίες αρκετών εικονοστοιχείων. Σε κάθε περίπτωση, το μοντέλο μας αποδίδει καλύτερα από οποιαδήποτε συγκρινόμενη μέθοδο έχοντας λιγότερες παραμέτρους εκπαίδευσης από τη δεύτερη καλύτερη (FFVI).

| **Method** | **PSNR**(↑) | **SSIM**(↑) | **MSE**(↓) | **# Parameters** |
|---|---|---|---|---|
| SpAGAN [13] | 33.03 | 0.9211 | 53.84 | 2.98M |
| Median | 32.89 | 0.8540 | 55.72 | NA |
| FFVI [9] | 44.61 | 0.9796 | 2.977 | 35.9M |
| CloudTran Parallel | 50.58 | 0.9891 | 3.098 | 3.2M |
| CloudTran | **54.09** | **0.9943** | **0.970** | |

**Πίνακας 5.7:** Σύγκριση της προτεινόμενης μεθόδου CloudTran με άλλες πρότυπες μεθόδους για αποτελέσματα διαστάσεων 64×64.

| **Method** | **PSNR**(↑) | **SSIM**(↑) | **MSE**(↓) | **# Parameters** |
|---|---|---|---|---|
| SpAGAN [13] | 30.88 | 0.8916 | 78.98 | 2.98M |
| Median | 32.46 | 0.8741 | 58.912 | NA |
| FFVI [9] | 48.31 | 0.9922 | 1.373 | 35.9M |
| CloudTran Parallel | 50.26 | 0.9935 | 2.024 | 12.7M |
| CloudTran | **51.34** | **0.9950** | **1.202** | |

**Πίνακας 5.8:** Σύγκριση της προτεινόμενης μεθόδου CloudTran με άλλες πρότυπες μεθόδους για αποτελέσματα διαστάσεων 256×256.

| **SEN12MS-CR-TS** | |
|---|---|
| **PSNR** (dB) | 50.00 |
| **SSIM** | 0.9931 |
| **MSE** | 6.426 |

**Πίνακας 5.9:** Αποτελέσματα της μεθόδου CloudTran στα δεδομένα ελέγχου του SEN12MS-CR-TS.

Στον Πίνακα 5.9 παρουσιάζονται οι μετρικές της μεθόδου CloudTran για τα δεδομένα ελέγχου του σετ δεδομένων SEN12MS-CR-TS. Για τον σκοπό αυτό εκπαιδεύεται ένα ολοκληρωμένο μοντέλο και με τα δύο δίκτυα, για 100000 επαναλήψεις. Οι αναγραφόμενες τιμές δεν είναι άμεσα συγκρίσιμες με αυτές που παρουσιάζονται στην δουλειά των [19] για πλήθος λόγων, μερικοί εκ των οποίων είναι η χρήση SAR απεικονίσεων (Sentinel-1) και όλου του φάσματος των Sentinel-2 δεδομένων (13 κανάλια) σε αντίθεση με την δική μας μέθοδο όπου αξιοποιούνται μόνο τρία

κανάλια του οπτικού φάσματος (R, G, B) και η διαφορά στον τρόπο αξιολόγησης των μοντέλων. Παρ'ολα αυτά, οι τιμές που καταγράφονται στον Πίνακα 5.9 είναι σαφώς βελτιωμένες συγκριτικά με αυτές που παρουσιάζουν οι [19].

Στη συνέχεια παρουσιάζονται αποτελέσματα, τυχαία επιλεγμένα, όπως προκύπτουν από τα δεδομένα αξιολόγησης (validation set) του ιδιόκτητου σετ δεδομένων και του SEN12MS-CR-TS. Η πρώτη απεικόνιση κάθε σελίδας παρουσιάζει τις ground truth εικόνες σε διαστάσεις $64\times64$ και η δεύτερη τις ίδιες εικόνες μετά την εφαρμογή της τυχαίας μάσκας συννέφων. Στην τρίτη παρουσιάζονται τα αντίστοιχα αποτελέσματα που προκύπτουν από το βασικό δίκτυο, δηλαδή οι απαλλαγμένες από σύννεφα απεικονίσεις, διαστάσεων $64\times64$. Οι επόμενες δύο εικόνες περιέχουν κατά σειρά, τα αποτελέσματα του spatial upsampler, δηλαδή την τελική αναδομημένη απεικόνιση του μοντέλου CloudTran και την ground truth εικόνα στις αρχικές διαστάσεις $256\times256$. Τέλος, παρουσιάζεται αθροιστικά ο αριθμός των έγκυρων εικονοστοιχείων (που δεν καλύπτονται από σύννεφα).

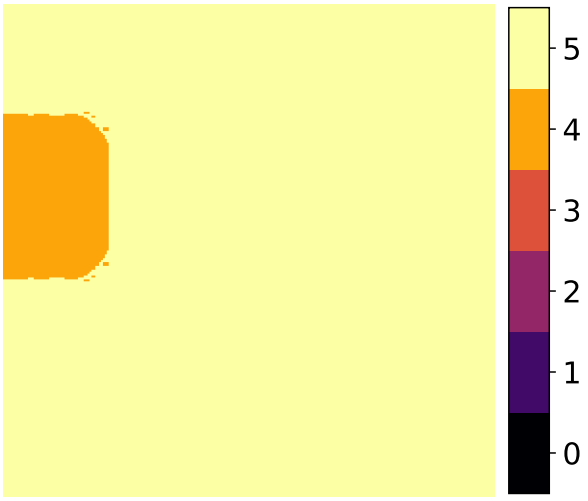
Downsampled GT

Masked Target

Core Output

Upsampler Output

GT




Coverage

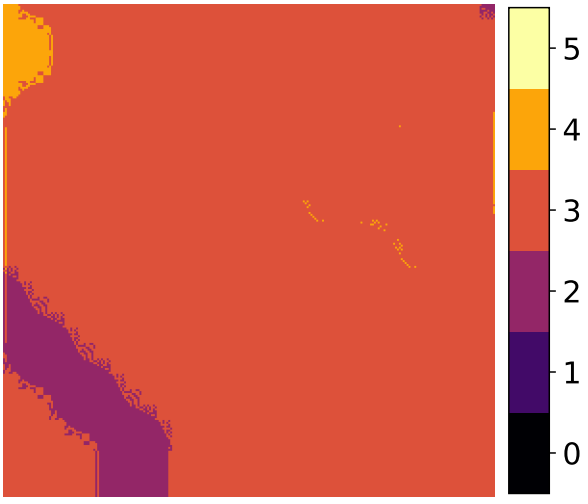
Downsampled GT

Masked Target

Core Output

Upsampler Output

GT

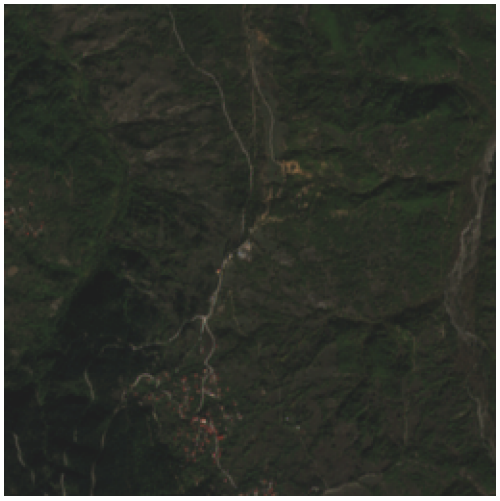



Coverage

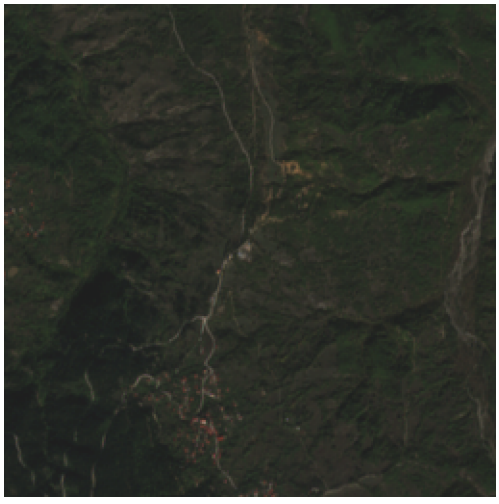
Downsampled GT

Masked Target

Core Output

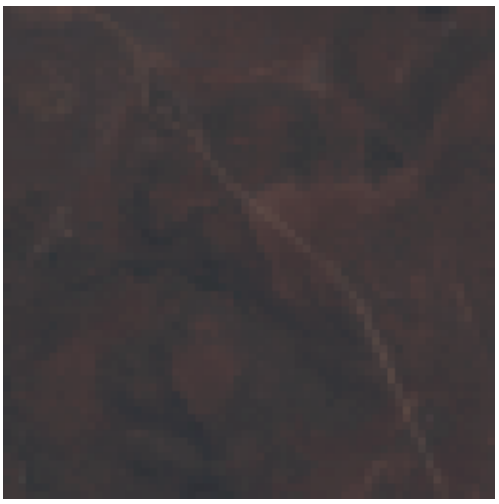


Upsampler Output

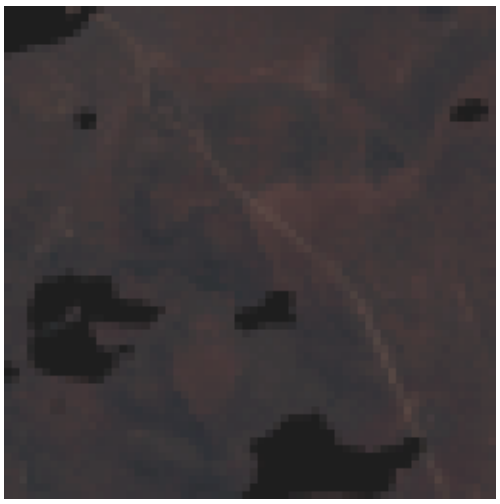


GT


Coverage

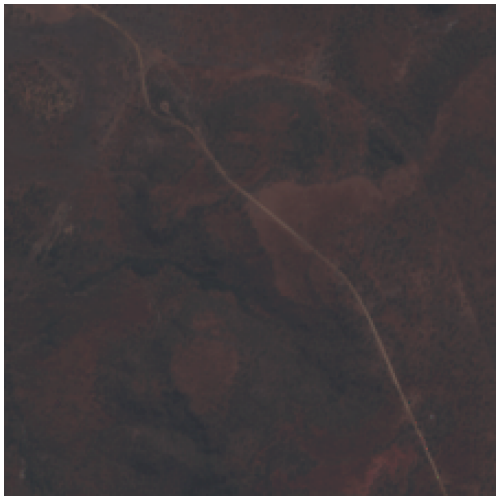


Downsampled GT

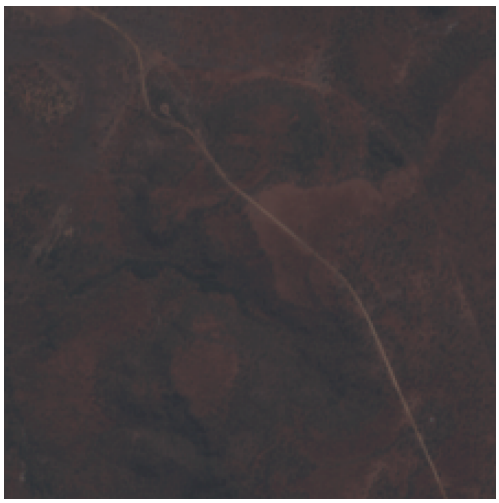


Masked Target

Core Output

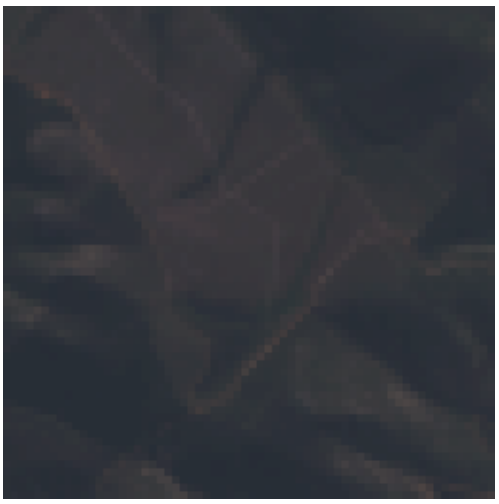


Upsampler Output



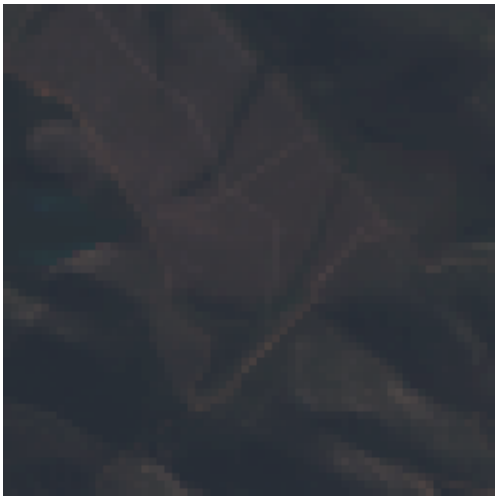
GT


Coverage

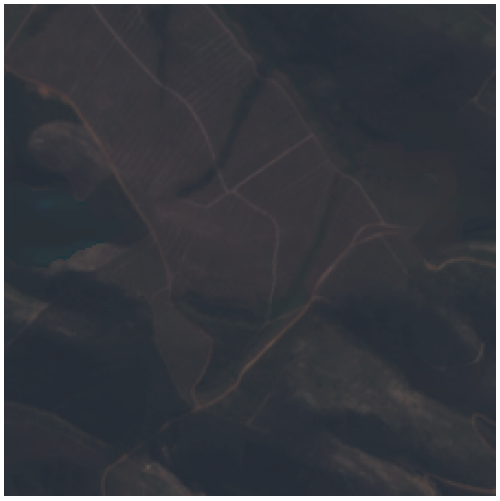


Downsampled GT

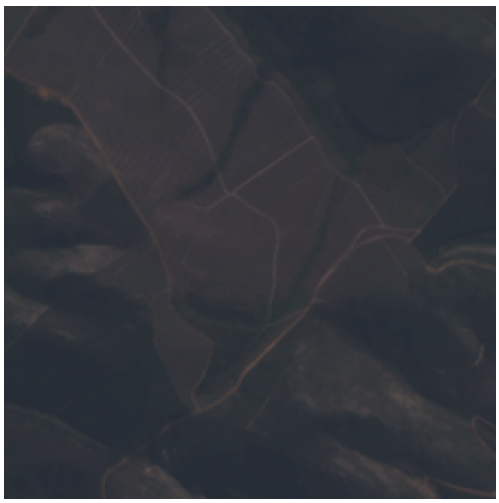
Masked Target



Core Output



Upsampler Output



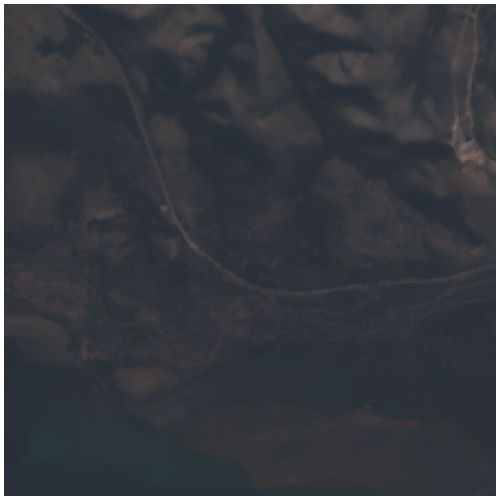
GT


Coverage

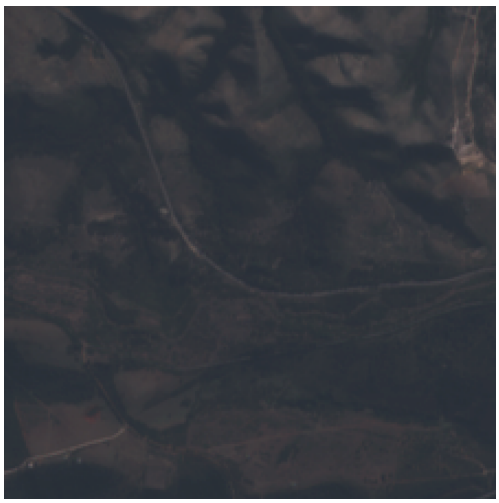
Downsampled GT

Masked Target

Core Output



Upsampler Output



GT


Coverage

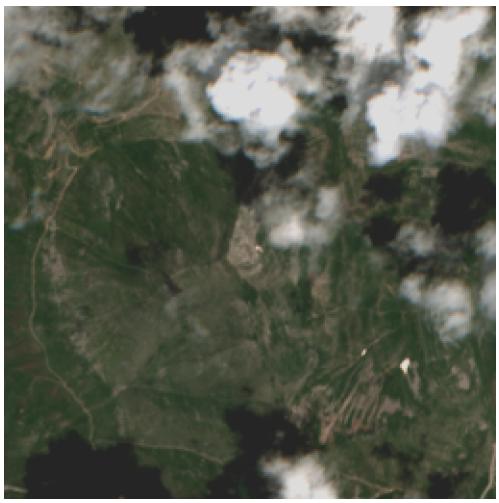
Προχωρόντας στις επόμενες σελίδες παρουσιάζονται αποτελέσματα που προκύπτουν από τα δεδομένα ελέγχου (test set) και των δύο σετ δεδομένων που έχουν προαναφερθεί, με την τελευταία εικόνα $I_T$ να είναι απαραίτητα επηρεασμένη από σύννεφα χωρίς κάποιον περιορισμό στο μέγιστο ποσοστό νεφοκάλυψης. Στη πρώτη σειρά παρουσιάζονται τα αποτελέσματα των δύο μεθόδων σύγκρισης, του φίλτρου ενδιάμεσης τιμής και του FFVI αντίστοιχα. Στη δεύτερη σειρά παρατίθενται τα αποτελέσματα της μεθόδου μας, πρώτα από τον Encoder και ύστερα από τον Decoder που αποτελεί και την πλήρη έκδοση της μεθόδου CloudTran. Όπως και προηγουμένως οι δύο τελευταίες εικόνες αποτελούνται από την αρχική, προς επεξεργασία, απεικόνιση επηρεασμένη από σύννεφα και από την διαθεσιμότητα της πληροφορίας σε επίπεδο εικονοστοιχείου.

Median

FFVI

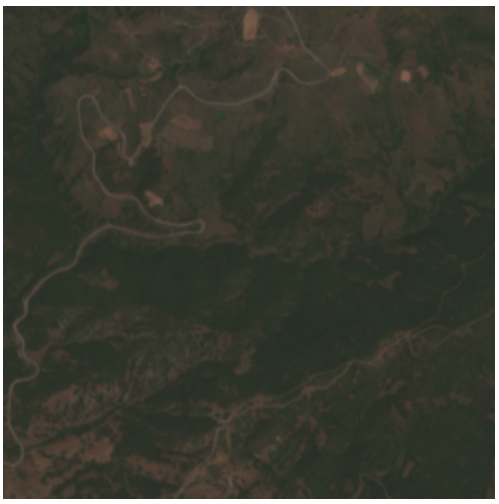
CloudTran Parralel

CloudTran

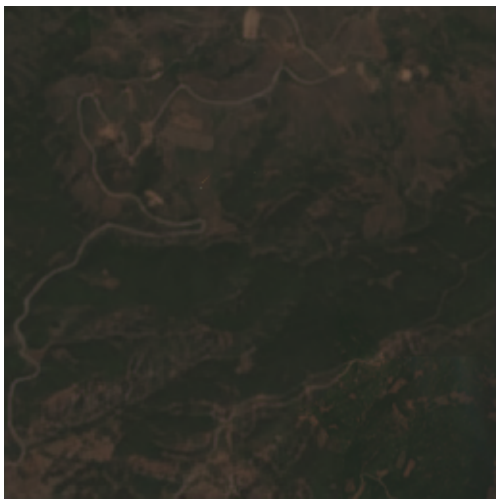


Target Image

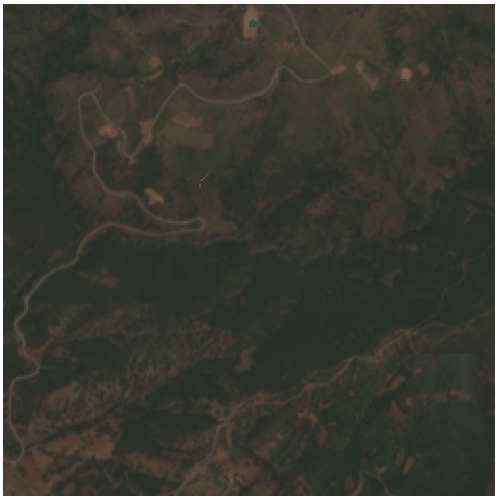

Coverage



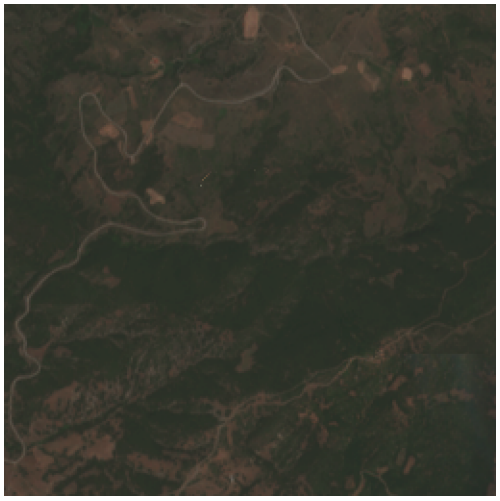
Median



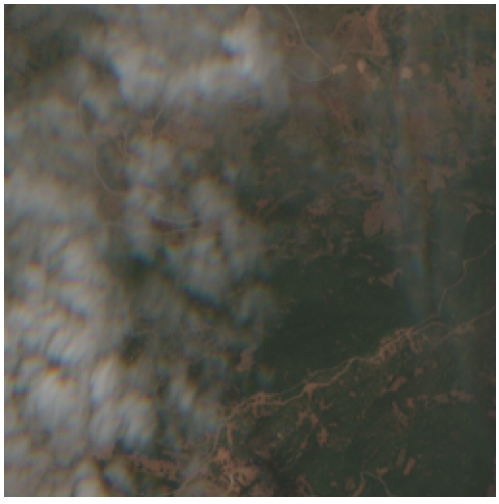
FFVI



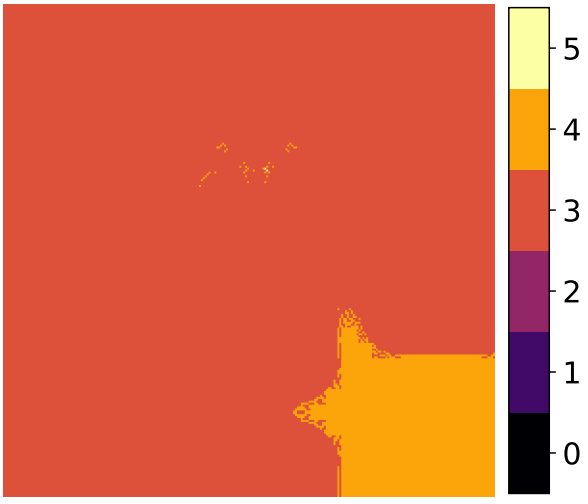
CloudTran Parralel



CloudTran



Target Image



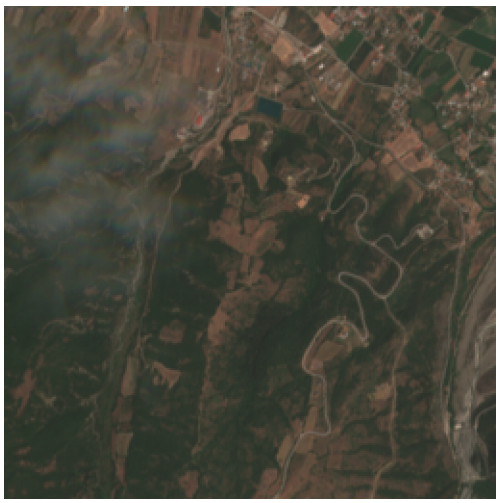

Coverage

Median

FFVI

CloudTran Parralel

CloudTran



Target Image


Coverage

Median

FFVI

CloudTran Parralel

CloudTran

Target Image


Coverage

Median

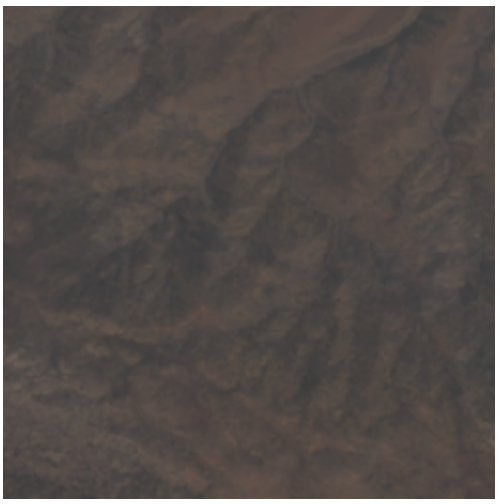
FFVI

CloudTran Parralel

CloudTran

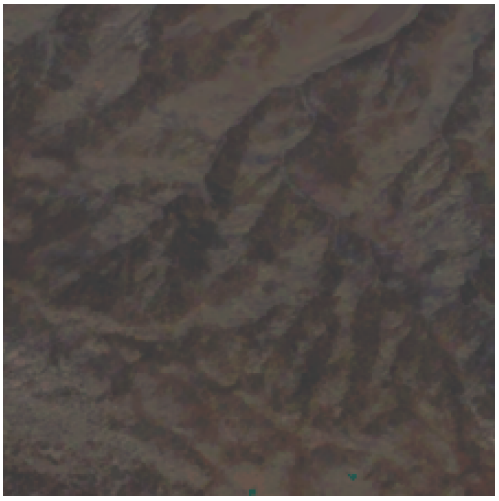
Target Image


Coverage

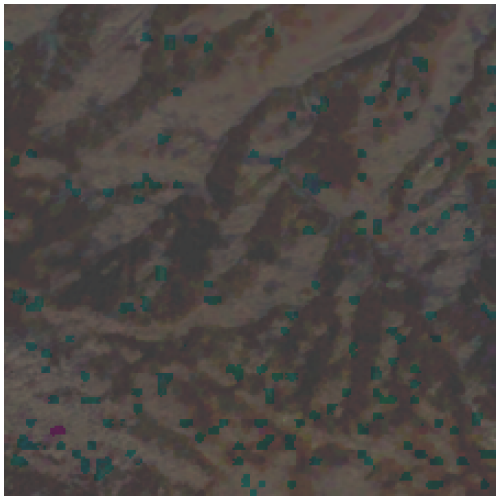


Median

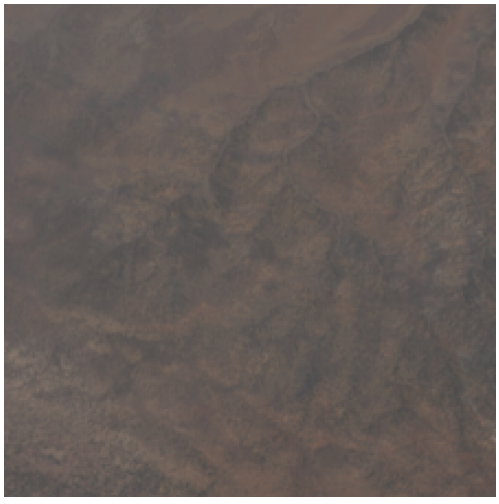
FFVI



CloudTran Parralel



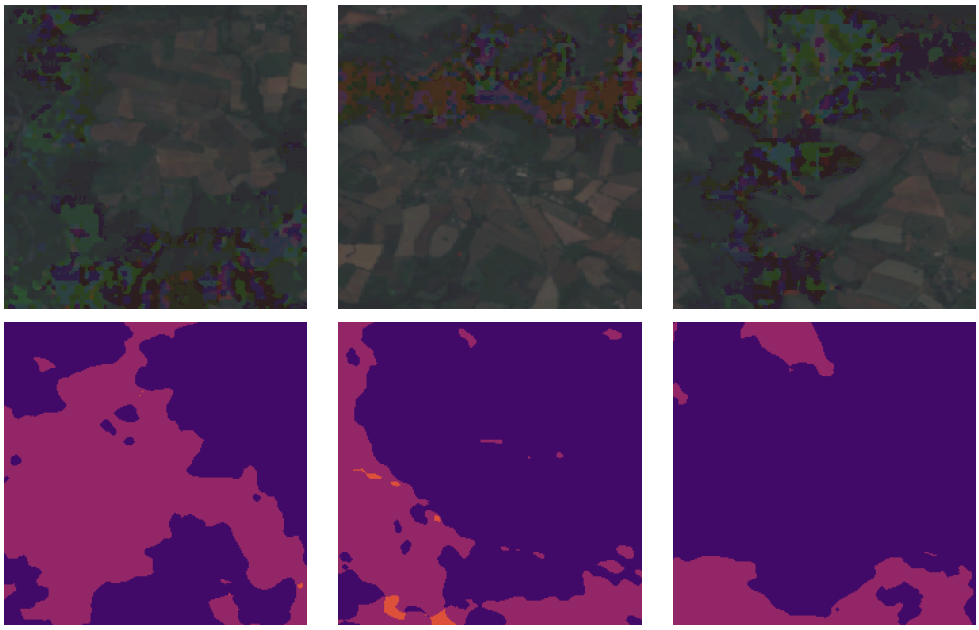
CloudTran



Target Image


Coverage

Τέλος, μερικές περιπτώσεις αστοχίας της μεθόδου παρουσιάζονται στο Σχήμα 5.42. Τα συγκεκριμένα αποτελέσματα οφείλονται στα πολύ μεγάλα ποσοστά νεφοκάλυψης (>50%) που παρατηρούνται σε όλες τις χρονικές στιγμές του κύβου $\mathcal{H}$ ο οποίος αποτελεί την είσοδο των δικτύων, εκτός από μια ή, σπανιότερα, δύο εξ αυτών (δεύτερη γραμμή του Σχήματος 5.42). Τέτοιου είδους αστοχίες μπορούν να αντιμετωπιστούν αυξάνοντας το μέγιστο ποσοστό νεφοκάλυψης κατά την διάρκεια της εκπαίδευσης και τον αριθμό των διαθέσιμων χρονικών στιγμών που απαρτίζουν τον κύβο $\mathcal{H}$.



**Σχήμα 5.42:** Περιπτώσεις αποτυχίας της μεθόδου CloudTran.

# ΚΕΦΑΛΑΙΟ 6

# Συμπεράσματα

## 6.1 Γενικά Συμπεράσματα

Στην παρούσα διπλωματική εργασία προτείνεται μια νέα μέθοδος αφαίρεσης συννέφων από χρονοσειρές δορυφορικών υπερφασματικών δεδομένων Sentinel-2, η αρχιτεκτονική της οποίας βασίζεται σε δύο δίκτυα δομής Transformer [6] με θεμέλιο λίθο τις συνδυαζόμενες στρώσεις axial attention [10]. Με τη χρήση του axial attention περιορίζουμε κατά πολύ το υπολογιστικό κόστος χωρίς να επηρεάζεται αρνητικά το οπτικό πεδίο του μοντέλου. Το βασικό δίκτυο ανακατασκευής (core) έχει τη μορφή Encoder-Decoder μοντέλου και αναλαμβάνει την αναδόμηση της εικόνας αναφοράς απαλλαγμένης από σύννεφα, στην εκάστοτε επιλεγμένη χαμηλή χωρική ανάλυση, τροφοδοτούμενο με μια χρονοσειρά υπερφασματικών δεδομένων στην οποία όλες οι περιοχές επηρεασμένες από σύννεφα έχουν μασκαριστεί αναλόγως. Το δίκτυο του spatial upsampler δομείται αποκλειστικά με Encoder και έχει ως στόχο την επαναφορά της παραγώμενης απεικόνισης, από το βασικό δίκτυο, στην αρχική χωρική διακριτική ικανότητα. Ο αλγόριθμος που προτείνεται, προσφέρει σημαντικά βελτιωμένα αποτελέσματα συγκριτικά με άλλες πρότυπες μεθόδους που αντιμετωπίζουν το ζήτημα αφαίρεσης συννέφων.

## 6.2 Ειδικά/Τεχνικά Συμπεράσματα

Η καινοτομία που προτείνεται σε αυτή την εργασία βασίζεται στην χρήση του axial attention ως τον βασικό μηχανισμό για την επεξεργασία πολυφασματικών δεδομένων όπως είναι οι δορυφορικές απεικονίσεις Sentinel. Τα πρότυπα δίκτυα Transformer επιβάλλουν πολύ υψηλό υπολογιστικό κόστος, τόσο κατά την διαδικασία της εκπαίδευσης όσο και κατά την παραγωγή των αποτελεσμάτων, ειδικά όταν αναφερόμαστε σε πολυφασματικές απεικονίσεις. Όπως αναφέραμε στην ενότητα 2.3 για μια τετραγωνική εικόνα διαστάσεων $N = S \times S$ το υπολογιστικό κόστος θα ήταν $\mathcal{O}(N^2)$. Με την χρήση και τον συνδυασμό πολλαπλών επιπέδων axial attention το υπολογιστικό κόστος, για την ίδια εικόνα, είναι $\mathcal{O}(N\sqrt{N})$, έχοντας πετύχει εξοικονόμηση βαθμού $\mathcal{O}(\sqrt{N})$. Παρά το γεγονός ότι με τον μηχανισμό αυτό κάθε άξονας της εικόνας επεξεργάζεται

ανεξάρτητα από τους υπόλοιπους, το οπτικό πεδίο του μοντέλου παραμένει καθολικό. Οι απαιτήσεις, σε χρόνο και σε πόρους, για την εκπαίδευση του μοντέλου δεν είναι πολύ μεγάλες, γεγονός που αποδικνύεται και από τον αριθμό των παραμέτρων που εισάγονται σε πειράματα πλήρους προδιαγραφών (3.2M αποκλειστικά για το βασικό δίκτυο και 12.7M για την πλήρη αρχιτεκτονική) συγκριτικά με άλλες πρότυπες μεθόδους όπως είναι το FFVI που εισάγει 35.9M παραμέτρους. Στους περιορισμούς της προτεινόμενης αρχιτεκτονικής, αντιθέτως, εντάσσεται η δειγματοληψία των αποτελεσμάτων, η οποία είναι πολύ χρονοβόρα παρά το γεγονός ότι χρησιμοποιείται μια ημι-παράλληλη αυτο-παλινδρομητική διαδικασία που δεν απαιτεί την εκ νέου εφαρμογή ολόκληρου του μοντέλου για την παραγωγή κάθε περιοχής.

## 6.3 Προοπτικές

Η παρούσα εργασία έχει δημοσιευτεί στο συνέδριο του ISPRS 2022 (International Society for Photogrammetry and Remote Sensing) [18] ενώ ήδη συλλογιζόμαστε και εργαζόμαστε σε πιθανές βελτιώσεις και επεκτάσεις της μεθόδου CloudTran. Πειράματα έχουν προγραμματιστεί να γίνουν με την χρήση επιπλέον καναλιών του δορυφόρου Sentinel-2 για την εκπαίδευση του μοντέλου. Ως τώρα χρησιμοποιούνταν αποκλειστικά συνθετικές RGB απεικονίσεις, ωστόσο γνωρίζουμε ότι τα κανάλια του υπέρυθρου φάσματος εμφανίζουν μικρές διαπεραστικές ιδιότητες στα λεπτά σύννεφα, γεγονός το οποίο μπορεί να ανεβάσει την ποιότητα του εκπαιδευόμενου μοντέλου. Μια ακόμα ιδέα προς υλοποίηση αφορά τον τρόπο αναδόμησης των νέων απεικονίσεων. Το κάθε μοντέλο εκπαιδεύεται ανεξάρτητα σε κάθε κανάλι των απεικονίσεων εισόδου μέσω κωδικοποίσης θέσης (positional encoding), ωστόσο ως ημι-παράλληλο αυτοπαλινδρομικό μοντέλο (το βασικό δίκτυο ανακατασεκυής) θα μπορούσε κατά την παραγωγή αποτελεσμάτων να επηρεάζεται και από τα υπόλοιπα κανάλια με σκοπό την αποφυγή περιπτώσεων θορύβου όπως συμβαίνει σε ορισμένες περιπτώσεις. Τέλος, με στόχο την επιβεβαίωση της ποιότητας και της αξιοπιστίας της μεθόδου CloudTran, προορίζεται η σύγκριση της με επιπλέον πρότυπες μεθόδους αφαίρεσης συννέφων όπως η [19] που δημοσιεύθηκε το 2022.

# Βιβλιογραφία


[1] Zhou Wang et al. «Image quality assessment: from error visibility to structural similarity». In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[2] Ian J. Goodfellow et al. «Generative Adversarial Networks». In: *arXiv preprint arXiv:1406.2661* (2014).

[3] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI*. 2015, pp. 234–241.

[5] Phillip Isola et al. «Image-To-Image Translation With Conditional Adversarial Networks». In: *IEEE. CVPR*. 2017.

[6] Ashish Vaswani et al. «Attention is all you need». In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[7] Praveer Singh and Nikos Komodakis. «Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks». In: *IGARSS*. IEEE. 2018, pp. 1772–1775.

[8] Lin Yan and David P Roy. «Large-area gap filling of Landsat reflectance time series by spectral-angle-mapper based spatio-temporal similarity (SAMSTS)». In: *Remote Sensing* 10.4 (2018), p. 609.

[9] Ya-Liang Chang et al. «Learnable gated temporal shift module for deep video inpainting». In: *arXiv preprint arXiv:1907.01131* (2019).

[10] Jonathan Ho et al. «Axial attention in multidimensional transformers». In: *arXiv preprint arXiv:1912.12180* (2019).

[11] Michael Schmitt et al. *SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion*. 2019.

[12] Andrea Meraner et al. «Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion». In: *ISPRS Journal of Photogrammetry and Remote Sensing* 166 (2020), pp. 333–346.

[13] Heng Pan. «Cloud Removal for Remote Sensing Imagery via Spatial Attention Generative Adversarial Network». In: *arXiv preprint arXiv:2009.13015* (2020).

[14] D.P. Roy and L. Yan. «Robust Landsat-based crop time series modelling». In: *Remote Sensing of Environment* 238 (2020), p. 110810.

[15] Vishnu Sarukkai et al. «Cloud Removal from Satellite Images using Spatiotemporal Generator Networks». In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020.

[16] Lin Yan and David P. Roy. «Spatially and temporally complete Landsat reflectance time series modelling: The fill-and-fit approach». In: *Remote Sensing of Environment* 241 (2020), p. 111718.

[17] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. «Colorization Transformer». In: *ICLR 2021*. 2021.

[18] D. Christopoulos, V. Ntouskos, and K. Karantzalos. «CLOUDTRAN: CLOUD REMOVAL FROM MULTITEMPORAL SATELLITE IMAGES USING AXIAL TRANSFORMER NETWORKS». In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B2-2022 (2022), pp. 1125–1132.

[19] Patrick Ebel et al. «SEN12MS-CR-TS: A Remote Sensing Data Set for Multi-modal Multi-temporal Cloud Removal». In: *IEEE Transactions on Geoscience and Remote Sensing* (2022).

# Κατάλογος Πινάκων

# Κατάλογος Σχημάτων