National Technical University of Athens
MSc in Data Science and Machine Learning

Generation of Synthetic Referring Expressions for Object Segmentation in Videos

MASTER THESIS

IOANNIS V. KAZAKOS

Supervisor : Konstantinos Karantzalos
Associate Professor at NTUA

Athens, November 2020

National Technical University of Athens
MSc in Data Science and Machine Learning

Generation of Synthetic Referring Expressions for Object Segmentation in Videos

MASTER THESIS

IOANNIS V. KAZAKOS

Supervisor : Konstantinos Karantzalos
Associate Professor at NTUA

Approved by the examining committee on the November 11, 2020.

Konstantinos Karantzalos Associate Professor at NTUA	Xavier Giró-i-Nieto Associate Professor at UPC	Giorgos Stamou Associate Professor at NTUA

Athens, November 2020

Ioannis V. Kazakos

MSc in Data Science and Machine Learning

Περίληψη

Η ενσωμάτωση της επεξεργασίας φυσικής γλώσσας στην όραση υπολογιστών έχει σημειώσει σημαντική πρόοδο τα τελευταία χρόνια χάρη στη συνεχή εξέλιξη της βαθιάς μηχανικής μάθησης. Ένα καινοτόμο πρόβλημα το οποίο συνδυάζει τη μηχανική όραση και την επεξεργασία φυσικής γλώσσας είναι αυτό της κατάτμησης αντικειμένων σε ακολουθίες εικόνων (βίντεο) με τη χρήση αναφορικών εκφράσεων, όπου μια πρόταση φυσικής γλώσσας καθορίζει ποιο αντικείμενο πρέπει να κατατμηθεί σε ένα βίντεο. Μια από τις μεγαλύτερες προκλήσεις αυτής της εφαρμογής είναι η έλλειψη συνόλων δεδομένων μεγάλης κλίμακας, εξαιτίας του υπερβολικά μεγάλου χρόνου και ανθρώπινης προσπάθειας που απαιτούνται για τη συλλογή τους. Επιπλέον, τα υπάρχοντα σύνολα δεδομένων υποφέρουν από ετικέτες κακής ποιότητας καθώς, σύμφωνα με μια σχετική ανάλυση, περίπου μία στις δέκα αναφορικές εκφράσεις που περιέχονται σε αυτά αποτυγχάνουν να περιγράψουν μοναδικά το αντικείμενο-στόχο.

Ο σκοπός της παρούσας μεταπτυχιακής εργασίας είναι να αντιμετωπίσει αυτές τις προκλήσεις προτείνοντας μια καινοτόμο μέθοδο για την παραγωγή συνθετικών αναφορικών εκφράσεων για μια εικόνα (ενός καρέ του βίντεο). Η μέθοδος αυτή παράγει συνθετικές αναφορικές εκφράσεις χρησιμοποιώντας μόνο τις ετικέτες αναφοράς των αντικειμένων μιας εικόνας ή ενός βίντεο, καθώς και τα χαρακτηριστικά τους, τα οποία ανιχνεύονται από ένα υπερσύγχρονο βαθύ νευρωνικό δίκτυο εκπαιδευμένο για τον εντοπισμό αντικειμένων και των χαρακτηριστικών τους. Ένα από τα πλεονεκτήματα της προτεινόμενης μεθόδου είναι ότι ο ορισμός της επιτρέπει την εφαρμογή της σε οποιοδήποτε άλλο σύνολο δεδομένων εντοπισμού ή κατάτμησης αντικειμένων.

Χρησιμοποιώντας την προτεινόμενη μέθοδο, δημιουργείται και παρουσιάζεται το πρώτο μεγάλης κλίμακας σύνολο συνθετικών δεδομένων με αναφορικές εκφράσεις για κατάτμηση αντικειμένων σε βίντεο, βασισμένο σε ένα υπάρχον σύνολο δεδομένων κατάτμησης αντικειμένων σε βίντεο. Η παρούσα εργασία περιλαμβάνει στατιστική ανάλυση καθώς και σύγκριση του παραγόμενου συνόλου συνθετικών δεδομένων με υπάρχοντα σύνολα δεδομένων κατασκευασμένα από τον άνθρωπο.

Τα πειράματα που διεξήχθησαν σε τρία διαφορετικά σύνολα δεδομένων που έχουν χρησιμοποιηθεί για την κατάτμηση αντικειμένων σε βίντεο με τη χρήση αναφορικών εκφράσεων, αποδεικνύουν την αποτελεσματικότητα των παραγόμενων συνθετικών δεδομένων. Συγκεκριμένα, τα αποτελέσματα επιδεικνύουν ότι προ-εκπαιδεύοντας ένα βαθύ νευρωνικό δίκτυο με το προτεινόμενο σύνολο συνθετικών δεδομένων, είναι δυνατή η βελτίωση της ικανότητας γενίκευσης του δικτύου σε διαφορετικά σύνολα δεδομένων. Το συγκεκριμένο αποτέλεσμα έχει ακόμα μεγαλύτερη αξία αν αναλογιστεί κανείς ότι η επίτευξή του δε συμπεριλαμβάνει κανένα επιπλέον κόστος για υποσημείωση δεδομένων από ανθρώπους.

Λέξεις κλειδιά

Όραση Υπολογιστών, Επεξεργασία Φυσικής Γλώσσας, Όραση και Γλώσσα, Αναφορικές Εκφράσεις, Κατάτμηση Αντικειμένων σε Βίντεο, Παραγωγή Συνθετικών Δεδομένων

Abstract

Integrating computer vision with natural language processing has achieved significant progress over the last years owing to the continuous evolution of deep learning. A novel vision and language task, which is tackled in the present Master thesis is referring video object segmentation, in which a language query defines which instance to segment from a video sequence. One of the biggest challenges for this task is the lack of relatively large annotated datasets since a tremendous amount of time and human effort is required for annotation. Moreover, existing datasets suffer from poor quality annotations in the sense that approximately one out of ten referring expressions fails to uniquely describe the target object, according to a relevant analysis.

The purpose of the present Master thesis is to address these challenges by proposing a novel method for generating synthetic referring expressions for an image (video frame). This method produces synthetic referring expressions by using only the ground-truth annotations of objects as well as their attributes, which are detected by a state-of-the-art object detection deep neural network. One of the advantages of the proposed method is that its formulation allows its application to any object detection or segmentation dataset.

By using the proposed method, the first large-scale dataset with synthetic referring expressions for video object segmentation is created, based on an existing large benchmark dataset for video instance segmentation. A statistical analysis and comparison of the created synthetic dataset with existing, human-produced datasets is also provided in the present Master thesis.

The conducted experiments on three different datasets used for referring video object segmentation prove the efficiency of the generated synthetic data. More specifically, the obtained results demonstrate that by pre-training a deep neural network with the proposed synthetic dataset one can improve the ability of the network to generalize across different datasets. This outcome is even more important taking into account that no additional annotation cost is involved.

Key words

Computer Vision, Natural Language Processing, Vision and Language, Referring Expressions, Video Object Segmentation, Synthetic Data Generation

Σύνοψη

Ο συνδυασμός της όρασης υπολογιστών και της επεξεργασίας φυσικής γλώσσας έχει προσελκύσει το ενδιαφέρον της επιστημονικής κοινότητας τα τελευταία χρόνια, μιας και θεωρείται ένα σημαντικό βήμα προς τη δημιουργία αυτόνομων συστημάτων τεχνητής νοημοσύνης τα οποία θα είναι ικανά να αξιοποιούν και τα δύο είδη πληροφορίας για την επίλυση προβλημάτων του πραγματικού κόσμου [Hu16a, Yu18, Ye19]. Ένα παράδειγμα τέτοιου προβλήματος, με το οποίο καταπιάνεται η παρούσα μεταπτυχιακή εργασία, αποτελεί η κατάτμηση αντικειμένων σε εικόνες και βίντεο με τη χρήση αναφορικών εκφράσεων φυσικής γλώσσας. Ως αναφορική έκφραση ορίζεται μια πρόταση φυσικής γλώσσας αν και μόνο αν αποτελεί ακριβή περιγραφή ενός συγκεκριμένου και κανενός άλλου αντικειμένου που εμφανίζεται στην ίδια σκηνή [Reit92]. Το πρόβλημα αυτό χρησιμοποιεί σαν οδηγό μια αναφορική έκφραση που περιγράφει ένα μοναδικό αντικείμενο-στόχο προκειμένου να το εντοπίσει σε επίπεδο εικονοστοιχείου διαχωρίζοντάς το από άλλα αντικείμενα του ίδιου ή άλλου τύπου.

Η πρόοδος αυτού του καινοτόμου ερευνητικού πεδίου έχει επωφεληθεί από την πρόσφατη πρόοδο της βαθιάς μηχανικής μάθησης η οποία για να είναι αποδοτική απαιτεί μεγάλο αριθμό δεδομένων. Ωστόσο, μια από τις κυριότερες προκλήσεις του προβλήματος με το οποίο ασχολείται η παρούσα μεταπτυχιακή εργασία είναι η έλλειψη μεγάλων συνόλων δεδομένων με βίντεο τα οποία να περιλαμβάνουν ταυτόχρονα ετικέτες αντικειμένων σε επίπεδο εικονοστοιχείου και εκφράσεις φυσικής γλώσσας, όπως είναι για παράδειγμα το RefCOCO [Kaze14] για στατικές εικόνες. Η δημιουργία ανάλογων συνόλων δεδομένων απαιτεί μεγάλη ποσότητα χρόνου και ανθρώπινης προσπάθειας, γεγονός που έχει ωθήσει την επιστημονική κοινότητα να επενδύσει σε μεθόδους όπως η ημι/αυτο-επιβλεπόμενη μάθηση και η χρήση συνθετικών δεδομένων. Τα συνθετικά δεδομένα έχουν χρησιμοποιηθεί αποτελεσματικά σε διάφορες ερευνητικές εργασίες τόσο στην όραση υπολογιστών σε προβλήματα όπως η εκτίμηση οπτικής ροής [Doso15], η ανίχνευση αντικειμένων [Peng15], η σημασιολογική κατάτμηση [Sale18] και η κατάτμηση αντικειμένων σε βίντεο [Khor19], όσο και σε προβλήματα που συνδυάζουν τη μηχανική όραση και την επεξεργασία φυσικής γλώσσας όπως η συλλογιστική μέσω εικόνων [Liu19] και η πλοήγηση μέσω όρασης και γλώσσας [Frie18].

Ακολουθώντας αυτή την κατεύθυνση, η παρούσα μεταπτυχιακή εργασία προτείνει μια καινοτόμο μέθοδο για την παραγωγή συνθετικών αναφορικών εκφράσεων για μια εικόνα (ενός καρέ του βίντεο), η οποία βασίζεται μόνο στις ετικέτες αναφοράς των αντικειμένων καθώς και στα χαρακτηριστικά τους, τα οποία ανιχνεύονται από ένα υπερσύγχρονο βαθύ νευρωνικό δίκτυο [Ren15] εκπαιδευμένο για τον εντοπισμό αντικειμένων και των χαρακτηριστικών τους. Πιο συγκεκριμένα, η προτεινόμενη μέθοδος παράγει συνθετικές αναφορικές εκφράσεις για ένα αντικείμενο-στόχο συνδυάζοντας τα χαρακτηριστικά του αντικειμένου που εντοπίζονται από το προαναφερθέν νευρωνικό δίκτυο με άλλες ιδιότητες όπως η κλάση του αντικειμένου, το σχετικό του μέγεθος και η σχετική του θέση με άλλα αντικείμενα, οι οποίες υπολογίζονται με βάση τις ετικέτες αναφοράς των αντικειμένων. Ο τρόπος με τον οποίο υπολογίζονται και συνδυάζονται οι εν λόγω ιδιότητες και τα χαρακτηριστικά προκειμένου να σχηματιστούν αναφορικές εκφράσεις περιγράφεται αναλυτικά και δίνονται ανάλογα παραδείγματα.

Η προτεινόμενη μέθοδος εφαρμόζεται σε ένα υπάρχον μεγάλης κλίμακας σύνολο δεδομένων κατάτμησης αντικειμένων σε βίντεο, το YouTube-VIS [Yang19], το οποίο χάρη σε αυτή τη μέθοδο εμπλουτίζεται με συνθετικές αναφορικές εκφράσεις, χωρίς κανένα κόστος που να αφορά ανθρώπινη εργασία. Ένα σημαντικό πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι ο ορισμός της επιτρέπει την εφαρμογή της σε οποιοδήποτε άλλο σύνολο δεδομένων εντοπισμού ή κατάτμησης αντικειμένων σε εικόνες ή βίντεο. Με την προτεινόμενη μέθοδο, είναι δυνατή η δημιουργία πολλαπλών αναφορικών εκφράσεων για το ίδιο αντικείμενο σε κάθε καρέ του βίντεο συνδυάζοντας τις διάφορες ιδιότητες και

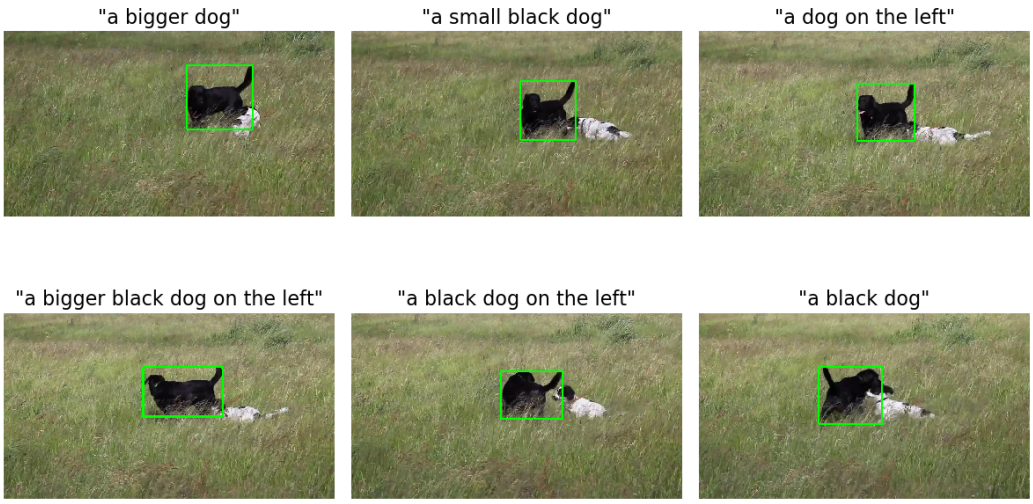


Figure 0.1: Example of synthetic referring expressions automatically generated with the proposed method. Multiple referring expressions can be created for the same video or even for the same frame.

τα χαρακτηριστικά του ώστε οι παραγόμενες συνθετικές εκφράσεις να το περιγράφουν μοναδικά. Ένα παράδειγμα διαφορετικών αναφορικών εκφράσεων οι οποίες παράγονται με την προτεινόμενη μέθοδο απεικονίζεται στην Εικόνα 0.1. Το παραγόμενο σύνολο δεδομένων, ονόματι SynthRef-YouTube-VIS, αποτελεί το πρώτο μεγάλης κλίμακας σύνολο συνθετικών αναφορικών εκφράσεων για κατάτμηση αντικειμένων σε βίντεο περιλαμβάνοντας 2,238 βίντεο και 15,798 διαφορετικές συνθετικές αναφορικές εκφράσεις. Η στατιστική ανάλυση και η σύγκριση του παραγόμενου συνόλου συνθετικών δεδομένων με υπάρχοντα σύνολα δεδομένων κατασκευασμένα από τον άνθρωπο δείχνει ότι αυτό υπερτερεί όσον αφορά το μέσο όρο διαφορετικών αναφορικών εκφράσεων ανά αντικείμενο.

Η προτεινόμενη μέθοδος αλλά και το εν λόγω σύνολο δεδομένων αξιολογούνται μέσω πειραμάτων τα οποία πραγματοποιούνται χρησιμοποιώντας ένα βαθύ νευρωνικό δίκτυο που ονομάζεται RefVOS [Bell20]. Το δίκτυο αυτό χρησιμοποιεί δύο υπερσύγχρονα μοντέλα για την κωδικοποίηση από τη μια των εικόνων (καρέ του βίντεο) και από την άλλη των αναφορικών εκφράσεων. Συγκεκριμένα, το DeepLabv3 [Chen17b] που έχει χρησιμοποιηθεί με μεγάλη επιτυχία στη σημασιολογική κατάτμηση, κωδικοποιεί την οπτική είσοδο, και το BERT [Devl19], ένα από τα πιο επιτυχημένα μοντέλα αναπαράστασης φυσικής γλώσσας, κωδικοποιεί τις αναφορικές εκφράσεις. Ο κατάλληλος συνδυασμός των εξαγόμενων οπτικών και γλωσσικών χαρακτηριστικών παράγει την τελική κατάτμηση του αντικειμένου-στόχου σε κάθε καρέ του βίντεο.

Τα πειράματα που πραγματοποιήθηκαν στην παρούσα μεταπτυχιακή εργασία είναι δύο ειδών:

1. Το πρώτο αφορά στην προ-εκπαίδευση του μοντέλου με τη χρήση πραγματικών (παραγόμενες από ανθρώπους), συνθετικών (παραγόμενες με την προτεινόμενη μέθοδο) ή συνδυασμό και των δύο τύπων αναφορικών εκφράσεων και την αξιολόγησή του σε δύο διαφορετικά σύνολα δεδομένων, το DAVIS-2017 [Khor18] και το A2D Sentences [Gavr18].

2. Το δεύτερο πείραμα αποσκοπεί στην απευθείας σύγκριση πραγματικών και συνθετικών αναφορικών εκφράσεων μέσω της εκπαίδευσης του μοντέλου στο ίδιο σύνολο δεδομένων βίντεο, αφενός με πραγματικές εκφράσεις και αφετέρου με συνθετικές, και της αξιολόγησής του στο ίδιο σύνολο πραγματικών δεδομένων.

Τα αποτελέσματα του πρώτου πειράματος επιδεικνύουν ότι προ-εκπαιδεύοντας ένα βαθύ νευρωνικό δίκτυο με το προτεινόμενο σύνολο συνθετικών δεδομένων, είναι εφικτή η βελτίωση της ικανότητας γενίκευσης του δικτύου σε διαφορετικά σύνολα δεδομένων, ειδικά στην περίπτωση που τα συνθετικά δεδομένα χρησιμοποιούνται σε συνδυασμό με πραγματικά. Επίσης, ακόμα μεγαλύτερη βελτίωση όσον αφορά την ακρίβεια κατάτμησης εντοπίζεται όταν το προ-εκπαιδευμένο μοντέλο χρησιμοποιεί-

ται σε ένα διαφορετικό σύνολο δεδομένων από αυτό στο οποίο έχει εκπαιδευτεί. Αυτό το αποτέλεσμα είναι σημαντικό γιατί σε πολλές εφαρμογές του πραγματικού κόσμου, τα μοντέλα μηχανικής μάθησης δεν έχουν τη δυνατότητα να εκπαιδεύονται στο τελικό σύνολο δεδομένων, αλλά βασίζονται σε μεγάλο βαθμό στην προ-εκπαίδευση.

Από την άλλη, τα αποτελέσματα της σύγκρισης μεταξύ πραγματικών και συνθετικών αναφορικών εκφράσεων, η οποία διεξάγεται στο δεύτερο πείραμα, οδηγούν στο συμπέρασμα ότι οι πραγματικές εκφράσεις, όντας πιο πλούσιες στην περιγραφή των αντικειμένων, οδηγούν σε μεγαλύτερη ακρίβεια κατάμησης. Ωστόσο, αν αναλογιστεί κανείς το μεγάλο κόστος για τη συλλογή των πραγματικών εκφράσεων και το αντίστοιχο μηδενικό για τη δημιουργία των συνθετικών, τα αποτελέσματα είναι συγκρίσιμα. Και για τα δύο πειράματα παρουσιάζονται τόσο αναλυτικοί πίνακες με ποσοτικά αποτελέσματα όσο και ποιοτικά αποτελέσματα της κατάτμησης των αντικειμένων σε διαφορετικές ακολουθίες εικόνων.

Επίσης, παρατίθεται μελέτη της επίδρασης της πληροφορίας που εμπεριέχεται στις συνθετικές αναφορικές εκφράσεις στην τελική ακρίβεια κατάτμησης. Τα αποτελέσματα αυτής δείχνουν ότι όσο περισσότερες ιδιότητες του αντικειμένου συμπεριλαμβάνονται στις συνθετικές αναφορικές εκφράσεις (όπως για παράδειγμα η σχετική του θέση ή το χρώμα του), τόσο βελτιώνεται η τελική ακρίβεια κατάτμησης. Μια άλλη μελέτη εστιάζει στο κατά πόσο το "πάγωμα" της εκπαίδευσης του κωδικοποιητή των αναφορικών εκφράσεων φυσικής γλώσσας (BERT), όταν το ήδη προ-εκπαιδευμένο μοντέλο εκπαιδεύεται με συνθετικές εκφράσεις, μπορεί να συμβάλλει στην καλύτερη μετέπειτα γενίκευσή του στο τελικό σύνολο δεδομένων. Η μελέτη αυτή δείχνει ότι, παρότι τα αποτελέσματα διαφέρουν ανάλογα με το τελικό σύνολο δεδομένων, η διαφορά στην ακρίβεια κατάτμησης είναι αμελητέα.

Τέλος, η παρούσα μεταπτυχιακή εργασία ενθαρρύνει την περαιτέρω επέκτασή της προτείνοντας μελλοντικές κατευθύνσεις έρευνας. Αυτές αφορούν πρώτον στην εφαρμογή της προτεινόμενης μεθόδου σε άλλα σύνολα δεδομένων εντοπισμού και κατάτμησης αντικειμένων σε εικόνες και βίντεο και δεύτερον στην ενίσχυση της προτεινόμενης μεθόδου για την παραγωγή πιο πλούσιων συνθετικών εκφράσεων με την εισαγωγή άλλων ιδιοτήτων όπως για παράδειγμα οι σχέσεις μεταξύ των αντικειμένων που εμφανίζονται.

Ευχαριστίες

Η παρούσα μεταπτυχιακή εργασία, με την οποία ολοκληρώνω το διατμηματικό μεταπτυχιακό πρόγραμμα σπουδών με τίτλο "Επιστήμη Δεδομένων και Μηχανική Μάθηση" του Εθνικού Μετσόβιου Πολυτεχνείου, δε θα μπορούσε να έχει έρθει εις πέρας χωρίς τη συνεργασία και τη συμπόρευσή μου με διάφορους ανθρώπους, οι οποίοι με βοήθησαν καθ' όλη τη διάρκεια αυτής.

Θα ήθελα κατ' αρχάς να ευχαριστήσω θερμά τον κ. Κωνσταντίνο Καράντζαλο, Αναπληρωτή Καθηγητή Ε.Μ.Π., για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα μεταπτυχιακή εργασία πάνω στο θέμα το οποίο με ενδιέφερε, για την επίβλεψη της εργασίας μου συνολικά, καθώς και για τη συγκατάθεσή του προκειμένου να πραγματοποιήσω ένα μέρος αυτής σε πανεπιστήμιο της επιλογής μου στο εξωτερικό. Κατόπιν, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον κ. Xavier Giró-i-Nieto, Αναπληρωτή Καθηγητή του Πολυτεχνικού Πανεπιστημίου Καταλονίας, για την τιμή που μου έκανε να γίνω μέλος του εργαστηρίου του και για την άκρως εποικοδομητική συνεργασία μας. Επίσης, ευχαριστώ τον κ. Γιώργο Στάμου, Αναπληρωτή Καθηγητή Ε.Μ.Π., ο οποίος συναίνεσε στη μετακίνησή μου στο πανεπιστήμιο του εξωτερικού, καθώς και για την τιμή που μου έκανε να είναι μέλος της επιτροπής εξέτασης της μεταπτυχιακής εργασίας μου.

Ιδιαίτερα θερμές ευχαριστίες θα ήθελα να αποδώσω στον κ. Carles Ventura, Επίκουρο Καθηγητή στο Ανοικτό Πανεπιστήμιο Καταλονίας, την κα Miriam Bellver, Υποψήφια Διδάκτωρ στο Barcelona Supercomputing Center, την κα Carina Silberer, Επίκουρη Καθηγήτρια του Πανεπιστημίου της Στουτγκάρδης και τον κ. Γιάννη Καλαντίδη, Μεταδιδακτορικό Ερευνητή, για την άψογη συνεργασία μας και την καθοριστική βοήθεια που μου παρείχαν σε όλα τα στάδια της μεταπτυχιακής εργασίας μου. Ανάλογες ευχαριστίες θα ήθελα να αποδώσω στον κ. Βαλσάμη Ντούσκο, Μεταδιδακτορικό Ερευνητή, για την εξίσου σημαντική βοήθεια που μου προσέφερε κατά τη συγγραφή της εργασίας μου.

Τελευταίο αλλά εξίσου σημαντικό, θα ήθελα να ευχαριστήσω τους φίλους μου και κυρίως την οικογένειά μου για την υποστήριξη που μου παρείχαν καθ' όλη τη διάρκεια υλοποίησης της μεταπτυχιακής αυτής εργασίας, και ιδιαίτερα τη μητέρα μου Γεωργία, τον πατέρα μου Βησσαρίωνα και την αδερφή μου Εύα, στους οποίους αφιερώνω την παρούσα εργασία.

Ioannis V. Kazakos,

Αθήνα, 11η Νοεμβρίου 2020

Η εργασία αυτή είναι επίσης διαθέσιμη ως Τεχνική Αναφορά CSD-SW-TR-42-17, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Τεχνολογίας Λογισμικού, Νοέμβριος 2020.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Acknowledgements

The present Master thesis, which marks my graduation from the interdepartmental Master's degree in "Data Science and Machine Learning" of the National Technical University of Athens, could not have been completed without the cooperation and collaboration with various people who helped me throughout it.

First of all, I would like to express my sincere gratitude to Dr. Konstantinos Karantzalos, Associate Professor at NTUA, for the opportunity he gave me to write the present Master thesis on a topic of my interest, for the supervision of my Master thesis and also for supporting my wish of doing part of the thesis abroad at a university of my choice. Then, I would like to say a special thank you to Dr. Xavier Giró-i-Nieto, Associate Professor at Universitat Politècnica de Catalunya, for our excellent and fruitful collaboration as well as for the honour of including me in his research team. Moreover, I would like to thank Dr. Giorgos Stamou, Associate Professor at NTUA, for consenting to my mobility abroad and for being a member of my Master thesis examination committee.

I would also like to express my gratitude and appreciation to Dr. Carles Ventura, Assistant Professor at Universitat Oberta de Catalunya, Ms. Miriam Bellver, PhD Candidate at Barcelona Supercomputing Center, Dr. Carina Silberer, Assistant Professor at University of Stuttgart, and Dr. Yannis Kalantidis, Postdoctoral Researcher, for our perfect collaboration and their overall guidance and help throughout my Master thesis. I am also grateful to Dr. Valsamis Ntouskos, Postdoctoral Researcher, for his valuable help during the writing of my Master thesis.

Last but not least, I would like to thank my friends and my family for their support during this challenging time, especially my mother Georgia, my father Vissarion and my sister Eva, to whom I dedicate this Master thesis.

Ioannis V. Kazakos,

Athens, November 11, 2020

This thesis is also available as Technical Report CSD-SW-TR-42-17, National Technical University of Athens, School of Electrical and Computer Engineering, Department of Computer Science, Software Engineering Laboratory, November 2020.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Contents

List of Figures

Chapter 1

Introduction

Inspired by the great success of deep learning in the fields of computer vision (CV) and natural language processing (NLP), the research community has invested in the integration of the aforementioned fields, by proposing several vision and language tasks and by trying to build models capable of combining visual and linguistic information effectively. A recently proposed vision and language task, addressed in the present Master thesis, is *referring video object segmentation* in which, given a linguistic phrase and a video, the goal is to generate a binary mask for the referred object in all the video frames where it is present. A visual description of the aforementioned task is provided in Figure 1.1.

1.1 Vision and Language Integration

Recent advancements in deep learning research has led the fields of computer vision and natural language processing see a significant progress in several tasks independently. This success has also increased the interest in solving challenges that combine visual and linguistic information, *i.e.* the integration of vision and language. Integrating vision and language is considered an important step towards the creation of powerful artificial intelligence (AI) systems that will be able to reason by processing multi-modal input.

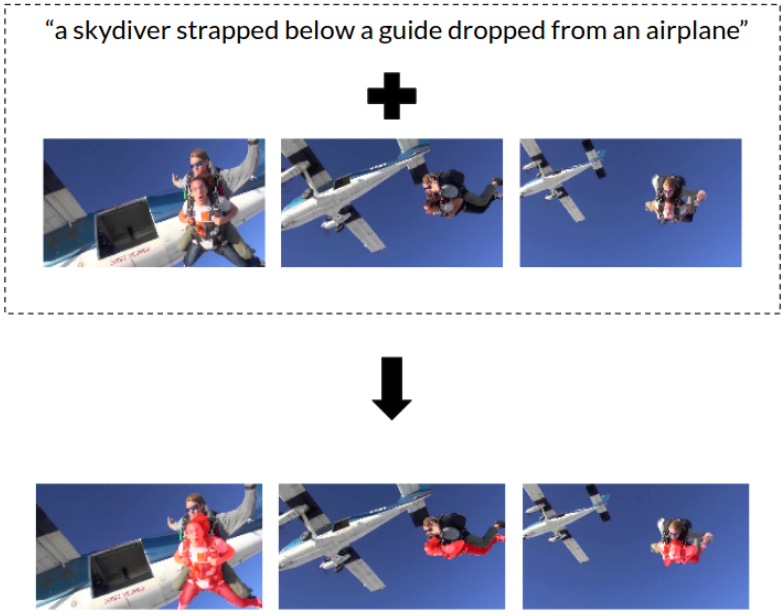


Figure 1.1: The task of referring video object segmentation. Top: A referring expression and a video are given as input. Bottom: A segmentation mask of the referent (highlighted in red) is produced at every frame. The provided referring expression is from the Refer-YouTube-VOS dataset [Seo20].

After more than half a century of research in neural networks and machine learning, deep learning has been establishing as the state-of-the-art technique of artificial intelligence since its breakthrough in 2012, when Krizhevsky et al. [Kriz12] presented a deep neural network, called AlexNet, which outperformed by a large margin all previous techniques in the Large Scale Visual Recognition Challenge (LSVRC). The release of large, high-quality, publicly available labelled datasets like ImageNet [Deng09], along with the empowerment of parallel GPU computing, which enabled the transition from CPU-based to GPU-based training, has led to the domination of deep learning in numerous AI fields, including computer vision and natural language processing.

More specifically, by using deep learning, computer vision has achieved prominent improvements in tasks such as visual content classification [Kriz12, Simo14, He16], object detection [Ren15, Redm16], semantic [Chen17a], instance [He17] and video object segmentation [Cael17, Vent19]. Convolutional neural networks (CNNs) [Fuku80, LeCu90] have become the standard approach for solving computer vision tasks. Most of the techniques rely on transferable general visual features by leveraging tasks such as image classification, detection, semantic segmentation, and action recognition. Usually, most preferred transferable global image representations are learned with deep CNN architectures like VGG [Simo14] and ResNets [He16] using large datasets such as ImageNet [Deng09]. These networks are used as the backbone of task-specific networks which transfer and enhance the obtained feature representations for solving downstream tasks.

Besides computer vision, deep learning has contributed to the significant progress in NLP research and its applications. For a long time, the majority of methods applied to NLP problems employed hand-crafted features using n-grams and bag-of-words [Joac98] models or standard machine learning techniques like Support Vector Machines (SVMs) [Cort95]. Such methods had been facing problems such as the curse of dimensionality since linguistic information was represented with high-dimensional features. However, with the recent popularity and success of word embeddings like word2vec [Miko13], which are low-dimensional, distributed representations, deep neural networks have achieved superior results on various language-related tasks as compared to previously used techniques.

Similar to CNNs for computer vision, several neural network architectures and techniques have been established in NLP research such as Recurrent Neural Networks (RNNs) [Rume86], Long Short-Term Memory (LSTM) [Hoch97] and the attention mechanism [Vasw17] in order to efficiently capture context in textual information. Especially in the last years, NLP has focused its efforts in solving multiple tasks at once with unsupervised pre-training of deep generalized language models like ELMo [Pete18], GPT-3 [Radf18] and BERT [Devl19], using large unlabeled corpora such as Wikipedia articles. These models have achieved incredible results in a wide variety of tasks such as machine translation, question answering and language inference.

Encouraged by the independent success of deep learning in CV and NLP fields, the research community has endeavored to build models combining vision and language. The aim of this integration is to produce systems which are able to provide complete understanding of visual and textual content at the same time. Several of the most important challenges that such systems have to tackle include:

- Generation of textual descriptions about visual content and vice versa, *i.e.* generation of visual content from textual descriptions
- Identification of objects and their relationships in visual content for reasoning or answering questions about them
- Navigation in an environment by leveraging input from both vision and natural language instructions
- Generation of short captions or longer stories about visual content
- Translation of textual content from one language to another using visual content for disambiguation

The aforementioned challenges can be associated to many practical applications of vision and language. One possible application in the biomedical domain can be the assistance of visually impaired individuals to get a holistic visual scene understanding by getting information about a scene from its textual descriptions and by answers received when asking questions about it. Other applications include automatic surveillance, autonomous driving, human-computer interaction and navigation.

Several tasks integrating language and vision have been proposed during the past years. An overview of them is depicted in Figure 1.2. These tasks include language observed in different levels such as words, phrases, sentences, paragraphs and documents while visual information is represented with images or videos. A brief description of the tasks presented in Figure 1.2 is provided below:

- *Referring Expression Generation and Comprehension/Segmentation*: Referring expression generation focuses on the creation of referring expressions (noun phrases) that identify specific entities called targets or referents [Mao16]. The inverse task is comprehension where target objects must be localized [Hu16b] or segmented [Hu16a] based on such expressions.

- *Visual Description Generation (Captioning)*: The goal of visual description generation or image/video captioning is to generate either global or dense descriptions of a given visual input in the form of a sentence [Elli13].

- *Visual Storytelling*: The aim of visual storytelling is to generate stories from one or more images or a video. Visual storytelling extends visual description generation by creating several sentences forming something similar to a paragraph [Huan16].

- *Visual Question Answering*: The goal of visual question answering (VQA) is to learn a model which comprehends the visual content at both global and local-level for finding an association with pairs of questions and answers in the natural language form [Anto15].

- *Visual Dialogue*: The goal of the visual dialogue task is to create an AI agent which, given an image, a history about dialogues and a question about the image, is able to infer context from the history, and answer the question accurately [Das17].

- *Visual Reasoning*: Visual reasoning targets to answer sophisticated queries by reasoning about the visual world. Efforts in this task have focused on creating diagnostic tests going beyond benchmarks such as VQA and reducing the biases of question-answer pairs by having detailed annotations describing the kind of reasoning each question requires [John17].

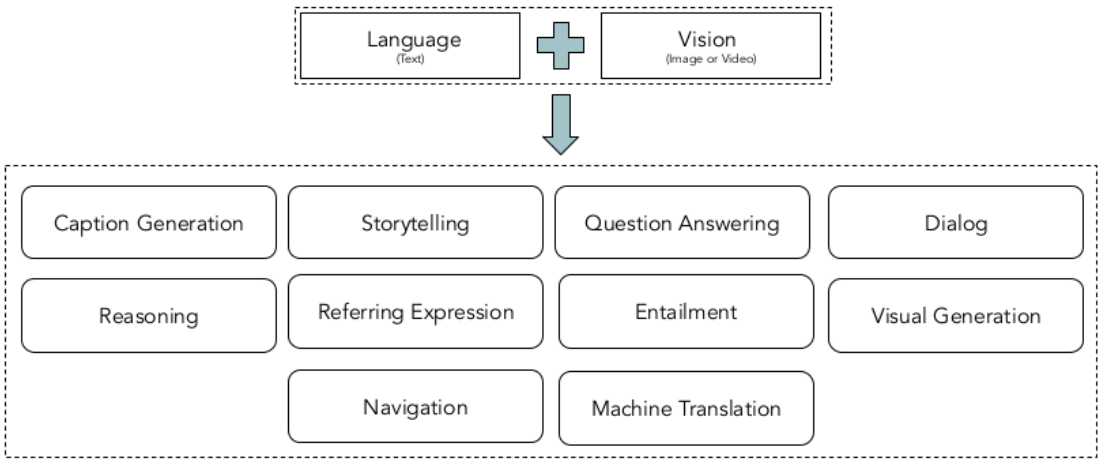


Figure 1.2: Different tasks combining vision and language [Moga19].

- *Image Entailment*: The task of predicting whether an image semantically entails a text, given image-sentence pairs where premise is defined by an image instead of a natural language sentence [Xie19].
- *Language-to-Vision Generation*: The aim of this task is to generate images/videos conditioned on natural language descriptions. The rapid evolution of generative adversarial networks (GANs) [Good14] has helped the growth of this task [Mans15].
- *Vision-and-Language Navigation*: The goal of vision-and-language navigation is to enable an agent or a robot to navigate within an environment given by the photo-realistic image views by interpreting natural language instructions [Ande18].
- *Multi-modal Machine Translation*: This task refers to the translation of a source language description into a target language using image/video information as additional context [Spec16].

1.2 Referring Video Object Segmentation

Video object segmentation is the task whose goal is to separate foreground objects from the background throughout a video sequence. This task has attracted wide attention lately due to its applicability to many practical problems including video analysis and video editing. Typically, this task has been addressed in semi-supervised or unsupervised setups. In the first case [Cael17, Vent19], a user manually annotates an object in a video frame and a system generates a pixel-wise binary mask for the object in the rest of the frames. The drawback of this setting is that pixel-wise annotations involve tedious and time-consuming human effort. In the unsupervised scenario [Goel18, Wang19b], estimation of object masks is performed without any guidance, by using salient features, independent motions, or known class labels. Although such approaches may be suitable for video analysis, according to Seo *et al.* [Seo20], the ambiguity and the lack of flexibility in defining foreground objects make them unsuitable for video editing which requires to segment arbitrary objects or their parts flexibly.

As an alternative approach, language referring expressions have been proposed as a different form of supervision for the task of video object segmentation. Referring expressions are linguistic expressions that allow the identification of an individual object (the *referent*) in a discourse or scene. According to computational linguistics and natural language processing community, a (noun) phrase is considered as a referring expression if it is an accurate description of the referent, but not of any other object in the current scene [Reit92]. Such linguistic expressions allow a more natural and direct human-computer interaction than interactive annotations in form of bounding boxes, masks, scribbles or points. Also, such expressions could be parsed from human speech processing systems allowing a direct human-machine communication in applications such as autonomous driving where the driver would refer to an object in the road scene and the car would identify it.

The task of video object segmentation using referring expressions is a novel task first addressed by Khoreva *et al.* [Khor18] in 2018 and later tackled in a similar setting from Gavrilyuk *et al.* [Gavr18] and Wang *et al.* [Wang19a] as "actor-action segmentation from a sentence". The name *referring video object segmentation*, in correspondence to referring image segmentation, was introduced by Seo *et al.* [Seo20] who also released the first large-scale benchmark for the task, called *Refer-YouTube-VOS*, in a concurrent work to the present Master thesis. An example illustrating the task of referring video object segmentation is provided in Figure 1.1.

1.3 Main Challenges and Motivation

Despite the increasing interest and research in the field, referring video object segmentation remains an extremely challenging task which is still far from being solved. The main challenges for the task which are presented below, are divided into those concerning data and those concerning models. In terms of models used for referring video object segmentation, challenges include:

- *Temporal consistency*: In contrast to static images, where the task of referring image segmentation has achieved significant progress, the video domain is more challenging as objects appearing in a video may disappear, reappear or be occluded from other objects. Consistency of segmentation masks across video frames is a challenging task and previous works have employed different techniques in order to achieve it. Recurrent architectures [Seo20] and 3D CNNs [Gavr18, Wang19a] are some of the ways previous works have tackled this challenge. Also, Khoreva *et al.* [Khor18] include a temporal consistency score in their objective function used for computing box proposals in each frame, based on the assumption that objects tend to move smoothly, and thus box proposals in consecutive frames should have a high overlap.

- *Model size*: Models used for the task of referring video object segmentation suffer from excessive size in terms of parameters which leads to huge memory requirements and the need of a great amount of time to train. The combination of vision and language demands at least two branches for encoding the visual and linguistic data, usually a deep CNN and a LSTM respectively. Especially in videos, where the time dimension is added, architectures can be even more complex in order to achieve temporal consistency as described above. Also, as presented in Chapter 2, many of these architectures use attention to effectively capture the dependencies between visual and linguistic features, adding extra layers and parameters to the deep architectures. This is a significant challenge considering the embedding of such models in cars or mobile phones which have restricted memory and computational resources.

The present Master thesis focuses on the challenges concerning the limitations of currently available datasets for the task of referring video object segmentation. In particular, the main challenges are:

- *Lack of large-scale datasets*: Before the release of Refer-YouTube-VOS [Seo20], which was created concurrently with the present Master thesis, no large-scale dataset existed for the task of referring video object segmentation. As it is explained in Subsection 2.2.2, existing datasets used for language-guided video object segmentation were limited either in terms of the number of videos [Khor18] or object classes [Gavr18]. Especially in deep learning and computer vision research, large datasets and benchmarks have proven their fundamental importance, enabling targeted progress and objective comparisons, thus their absence can impend the evolution of a scientific field. The annotation cost in terms of money and/or time is one of the main reasons for the absence of large-scale datasets. The present work addresses this challenge by proposing a method to automatically generate synthetic referring expression, thus eliminating human labour-intensive annotations.

Figure 1.3: Example of an invalid referring expression from A2D Sentences [Gavr18] dataset. The expression on top of the video frame fails to uniquely identify a specific object.

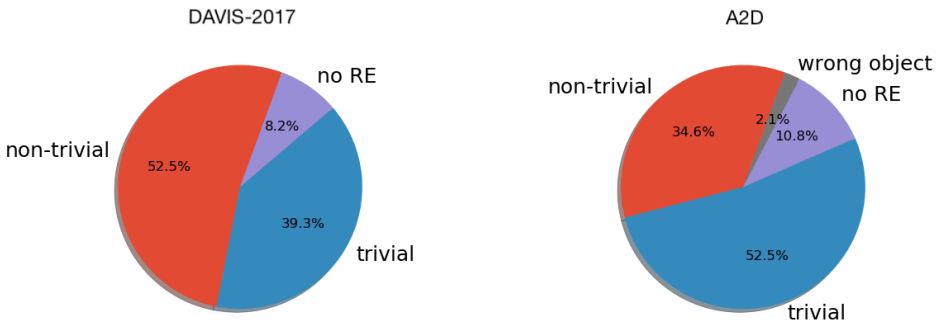


Figure 1.4: Categorization of referring expressions by their difficulty and correctness in the validation set of DAVIS-2017 and the test set of A2D Sentences [Bell20].

- *Referring expressions quality*: A previous work [Bell20] has argued that existing benchmarks used for video object segmentation with referring expressions are annotated with expressions that sometimes fail to address the objective of the task, namely to unambiguously refer to a specific object. Actually, because of the huge annotation time of large-scale dataset, language expressions in existing benchmarks for referring image and video object segmentation are mainly collected through crowdsourcing platforms like Amazon Mechanical Turk[1]. Although the crowdsourced annotations are usually validated from other experts, cases of bad annotations are still observed, like in the example of A2D Sentences [Gavr18], appearing in Figure 1.3. In this example, the provided referring expression (*"the man in red is running in a race"*) fails to unambiguously refer to a specific object, as it could be referring to three different instances present in the depicted video frame. An extensive analysis of the aforementioned work, illustrated in Figure 1.4, shows that approximately 10% of the referring expressions of two existing datasets used for the addressed task fail to uniquely describe the target object ("no RE" standing for "Not a Referring Expression").

As it is also observed in Figure 1.4, a significant proportion of the videos in existing datasets concerns trivial cases in which the target object could be identified with simple phrases. For example, in a video including one *dog* and one *ball*, each of them can be referred unambiguously, using just the class or supercategory in which it belongs *i.e.* simply saying *"a dog/animal"* and *"a ball"*.

Inspired by this simple scenario and the fact that existing datasets for object detection/segmentation are labeled in terms of the objects class, the idea of the present work is to create high quality synthetic referring expressions, starting from the referent's class and then enhancing them with other cues, without any human annotation cost. The proposed synthetic referring expressions are created on top of the YouTube-VIS [Yang19] dataset, which is described in detail in Section 3.1. The main advantage of this dataset is that all instances of a specified set of classes are annotated, allowing thus the creation of valid referring expressions.

1.4 Thesis Objectives and Structure

The present Master's thesis has the following objectives:

1. Study current methods in referring video object segmentation by reviewing related literature.
2. Underline the main challenges encountered on this task.
3. Propose a novel method for generating synthetic referring expressions.

[1] https://www.mturk.com/

4. Present and disseminate the first large-scale synthetic dataset for referring video object segmentation.

5. Evaluate the effectiveness of the proposed synthetic data in pre-training a deep neural network for the current task.

6. Compare the obtained performance using synthetic data with previous works.

7. Suggest future research directions regarding the use of synthetic referring expressions for video object segmentation.

Relevant literature is reviewed in *Chapter 2* where previous research works in the task of referring image and video object segmentation are explored by analyzing different techniques employed for solving the task. A review of object detection models based on deep learning is also included emphasizing on the one which is used in the proposed method. In the last section of this chapter, examples of scientific works using synthetic data in computer vision, natural language processing and their combination are presented.

Following, *Chapter 3* describes in detail the proposed method for generating synthetic referring expressions by explaining how different cues are combined for their creation. Moreover, Chapter 3 introduces the synthetic dataset created using the aforementioned method and includes an analysis of its statistics as well as some examples of synthetic referring expressions with their corresponding video frames.

Chapter 4 consists of an extensive analysis of the conducted experiments and the obtained results. More specifically, the training setups and quantitative evaluation metrics used in the experiments are described in detail and tables as well as figures of results comparing with previous works are illustrated.

Finally, in Chapter 5, the conclusions of the present thesis are summarized and future research directions for the topic under study are suggested.

Chapter 2

Literature Review

2.1 Referring Image Segmentation

2.1.1 Methods

Referring image segmentation, the task of segmenting objects or regions in images given a linguistic expression, was introduced by Hu *et al.* [Hu16a]. The authors distinguish this task from previous ones that were restricted to a fixed set of classes, like semantic segmentation [Long15, Chen17a], the task of predicting pixel-wise labels for a predefined set of object or stuff categories, or instance segmentation [He17], which additionally distinguishes different instances of an object class. Previous works about grounding natural language expressions were limited to only resolving a bounding box in an image [Hu16b, Mao16], therefore this was the first attempt of grounding language at pixel level.

The model they employ for solving this novel tasks consists of four main components which are depicted in Figure 2.1. The first is a language encoder based on a LSTM network. The input language expression is first converted into a sequence of fixed-length vectors using an embedding matrix. Then each of the τ word embeddings of the sequence $S = (w_1, ..., w_\tau)$ is processed by the LSTM network at each time step t. At the final time step $t = T$, when the the whole text sequence is processed by the LSTM, the hidden state h_τ of dimension $D_{text} = 1000$ is used as the encoded vector representation of the language expression. The second and third components of the model are two fully convolutional neural networks where the first of them is used as the image encoder and the second as a pixel classification network. The image encoder is a fully convolutional network as the one proposed by Long *et al.* [Long15] for semantic segmentation which, given an image of input $W \times X$, outputs a spatial feature map of dimension $w \times x \times D_{im}$. This means that the final spatial feature map includes $D_{im} = 1000$ local descriptors for each pixel of the pooled $w \times x$ image where

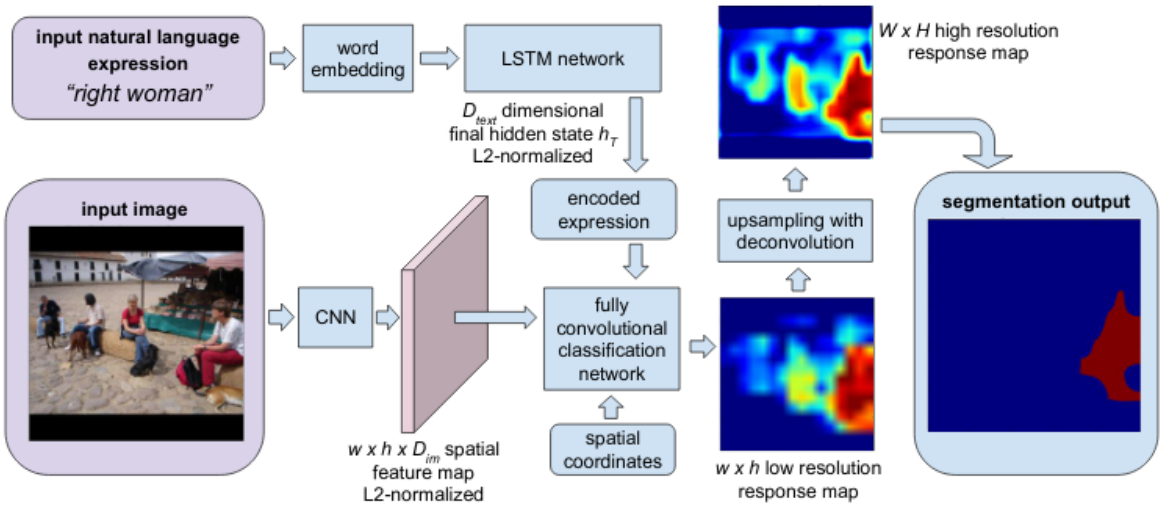


Figure 2.1: Model used by Hu *et al.* [Hu16a] who introduced the task of referring image segmentation.

$w = W/s$ and $h = H/s$ with $s = 32$. Also, two extra channels are added to each local descriptor, representing the relative coordinates of each pixel location from the upper left corner and the lower right corner of the feature map (represented as $(-1, -1)$ and $(+1, +1)$ respectively), so that the model can reason about spatial relationships found in expressions *e.g. "cat on the left"*.

After having extracted the visual and language features, h_τ is first tiled and concatenated to the local descriptor at each spatial location of the spatial feature map to obtain a $w \times h \times D^*$ (where $D^* = D_{im} + D_{text} + 2$) multi-modal spatial map containing both visual and linguistic features. The combined features are then passed through the third component of the model which is a two-layer fully convolutional classifier consisting of two 1×1 convolutions, that outputs a $w \times h$ low resolution segmentation map of the image. In order to recover the original image size, an upsampling operation using deconvolution (or transpose convolution) [Zeil11] is performed producing a $W \times H$ high resolution response map, whose values represent the confidence of whether a pixel belongs to the referred object. Since their work was the first to directly predict segmentation based on natural language expressions, the authors compare their model's performance with strong baselines they created such as segmentation from bounding boxes or classification proposals and combination of per-word semantic segmentation, indicating that their method outperforms all these baselines.

Instead of modeling the image and langauge features independently and then combining them to produce the segmentation map, subsequent works tried to jointly model the two modalities, in order to better exploit the correlations between words and image regions. Liu *et al.* [Liu17] combine visual and word features with an LSTM to recurrently refine the segmentation masks. Dynamic filters were used in [Marg18] and [Chen19b] to capture the recursive nature of language and the spatial information of the target object respectively. Li *et al.* [Li18] presented a recurrent refinement network (RRN) which refines the segmentation result by utilizing the feature pyramid structures in order to take advantage of multi-scale semantics.

Other works in referring image segmentation leverage attention to model the visual information of each word. MAttNet [Yu18] decomposes referring expressions using three modules related to the object's appearance, location and relationships with other objects and then uses both language and visual attention to direct each module to focus on the desired part of the expression and the image. Shi *et al.* [Shi18] use attention to extract keywords from a referring expression which are important for identifying the target object. Cross-modal self-attention is used in CMSA [Ye19] to better capture the long-range dependencies between linguistic and visual features. While STEP [Chen19a] works in the same direction, it also uses a convRNN [Xing15] to refine the textual representation and improve the segmentation. A recent work by Hu *et al.* [Hu20] proposes a bi-directional cross-modal attention module to learn the relationships between multi-modal features. Finally, Huang *et al.* [Huan20] use multi-modal graph reasoning to identify the correct object as well as suppress other irrelevant ones.

2.1.2 Relevant datasets

RefCOCO

RefCOCO is a large-scale dataset and benchmark for referring image segmentation. It is collected on top of the Microsoft COCO (Common Objects in Context) image collection [Lin14], which includes images of complex everyday scenes containing common objects in their natural context.

It is one of the three most frequently used benchmarks for referring image segmentation: RefCOCO, RefCOCO+ and RefCOCOg [Yu16]. RefCOCOg was collected using Amazon Mechanical Turk in a non-interactive setup, while RefCOCO and RefCOCO+ were collected using the Refer-it Game [Kaze14]. In this two-player game, the first player is shown an image with a segmented target object and asked to write a natural language expression referring to the target object. The second player is shown only the image and the referring expression and asked to click on the corresponding object. If the target object is correctly identified, the players receive points and swap roles. Otherwise, a new image and target object is assigned to them. Images in these collections were selected with the requirement to contain two or more objects of the same object category.

RefCOCO consists of 142,209 referring expressions for 50,000 objects in 19,994 images. The average number of words in its sentences is 3.61. Unlike RefCOCO+, where annotators are disallowed to use location words in their referring expressions, RefCOCO does not have any restrictions on its expressions. Moreover RefCOCO's referring expressions tend to be more concise than the ones from RefCOCOg which have an average length of 8.43. Another advantage of RefCOCO over RefCOCOg is that is contains more instances of same-category objects, having an average of 3.9 over 1.6 respectively.

Besides images, RefCOCO has been used for pre-training frame-based models on the task of referring video object segmentation like in the works of Khoreva *et al.* [Khor18] and Bellver *et al.* [Bell20], since a similar large-scale dataset for videos was not available. In the experiments of the present work, RefCOCO is also used along with the proposed synthetic dataset in order to assess how the synthetic data can contribute to a better pre-training of a deep neural network.

2.2 Referring Video Object Segmentation

2.2.1 Methods

Despite the increasing interest in referring image segmentation, only a few works have explored the segmentation of objects using referring expressions in the video domain *i.e.* referring video object segmentation. Khoreva *et al.* [Khor18] were the first to transfer the referring expression segmentation task from images to videos by collecting referring expressions for the DAVIS-2017 dataset [Pont17]. They use the image-based MAttNet [Yu18] model, pretrained on RefCOCO [Kaze14], to localize the target object, and then train a segmentation network with DAVIS-2017 to produce the pixel-wise prediction. They also employ a temporal consistency score in their objective function used for computing box proposals in each frame, in order to ensure a high overlap between box proposals in consecutive frames, based on the assumption that objects tend to move smoothly. Gavrilyuk *et al.* [Gavr18], in a relevant work, provide natural language sentences for Actor-Action Dataset (A2D) [Xu15] and J-HMDB [Jhua13] which are datasets used for action and human pose recognition and segmentation. They employ a 3D fully-convolutional model with dynamic filters in order to segment an actor in each frame of a video as specified by a language query. Although the task is similar to referring video object segmentation, the referring expressions they provide are intended to describe an actor and its action. The first large-scale dataset for referring video object segmentation, called Refer-YouTube-VOS, has been created concurrently to the present Master thesis by Seo *et al.* [Seo20] on top of YouTube-VOS [Xu18], a popular benchmark for video object segmentation. Besides the dataset, the authors propose a model called URVOS, which performs language-based object segmentation and mask propagation jointly using a single deep neural network. The network combines a cross-modal attention module, inspired by CMSA [Ye19] and a memory attention module to encourage temporal consistency across frames.

RefVOS

In another recent work, RefVOS [Bell20] has been the first model to leverage BERT [Devl19] for encoding the referring expressions. They have shown that using BERT instead of a bidirectional LSTM fed with GloVe embeddings [Penn14], which is a common practice in related works, brings significant improvements to the final segmentation. In the present Master thesis, RefVOS is the model used for the conducted experiments which aim at evaluating the proposed method and the generated synthetic dataset. A visual description of the architecture of RefVOS is depicted in Figure 2.3.

RefVOS is a frame-based model which uses DeepLabv3 [Chen17b] as its visual encoder. Convolutional Neural Networks deployed in fully convolutional fashion have shown to be effective for the task of semantic segmentation. However, the repeated combination of max-pooling and striding at consecutive layers of these networks significantly reduces the spatial resolution of the resulting feature maps. In order to recover the spatial resolution, deconvolutional (or transposed convolution)

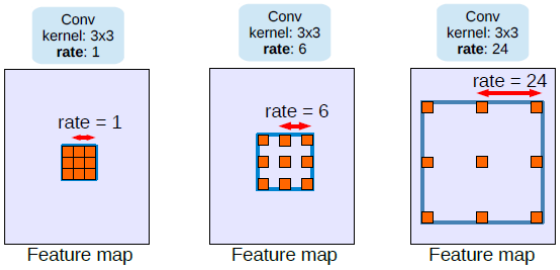


Figure 2.2: Illustration of atrous convolution with rates 1 (standard convolution), 6 and 24.

layers [Zeil11] have been employed in previous works using also skip connections to combine high resolution features from the contracting path to the upsampled output [Long15, Noh15, Ronn15].

Instead of decreasing and then increasing the feature maps spatial resolution, DeepLabv3 uses "atrous convolution", which was originally developed for the efficient computation of the undecimated wavelet transform in the "algorithme à trous" scheme [Hols90] and then used in the convolutional neural networks context [Gius13, Serm13, Papa15]. Considering two-dimensional signals, for each location i on the output y and a filter w, atrous convolution is applied over the input feature map x as:

$$y[i] = \sum_{k} x[i + r \cdot k]w[k] \tag{2.1}$$

where atrous rate r corresponds to the stride with which the input signal is sampled. This is equivalent to convolving the input x with upsampled filters produced by inserting $r - 1$ zeros between two consecutive filter values along each spatial dimension (hence the name "atrous convolution" where the French word "trous" means holes in English). Typical convolution is a special case of Equation 2.1 where $r = 1$. A visualization of atrous convolution with different atrous rates can be seen in Figure 2.2. By increasing the atrous rate r, one is able to use a wider field-of-view without the need to apply multiple convolutions or use larger kernels, *i.e.* without increasing the computational cost.

As seen in Figure 2.3 (top branch), DeepLabv3 applies four parallel atrous convolutions with different atrous rates, an architecture called Atrous Spatial Pyramid Pooling (ASPP) initially proposed in the first version of DeepLab [Chen17a], which is used in order to effectively capture multi-scale information. Besides the three 3×3 atrous convolutions, a 1×1 convolution and a global average pooling layer are involved. The features extracted from the five different operations are further processed in separate branches and fused to generate the final result. RefVOS model applies the ASPP architecture with atrous rates of 12, 24 and 36, as depicted in Figure 2.3.

The authors of DeepLabv3 also introduce the term of *output stride* to denote the ratio of input image resolution to the final feature map output resolution. Typical CNN architectures used for classification have an output stride of 32, meaning that the dimension of final feature responses, before fully connected layers, is 32 times smaller than the respective of the input image. Atrous convolutions, from the other side, allow to extract dense features without significantly decreasing the spatial resolution. RefVOS [Bell20] uses the architecture of DeepLabv3 with an output stride of eight.

Finally, in order to recover feature maps to the original image resolution for efficient segmentation, DeepLabv3 uses bilinear interpolation, which is sufficient in this setting because the feature maps produced with atrous convolutions are quite smooth. This way there is no need for extra deconvolutional (transpose convolution) layers which would increase the number of parameters and consequently memory requirements and total training time.

In order to obtain a linguistic embedding for the referring expression, RefVOS uses BERT, which stands for Bidirectional Encoder Representations from Transformers, and is a state-of-the-art language representation model presented by Devlin *et al.* [Devl19] (Google AI). Language model pre-training has been shown to be effective for improving performance in several natural language processing tasks. Before the publication of BERT, two typical strategies for applying pre-trained language rep-

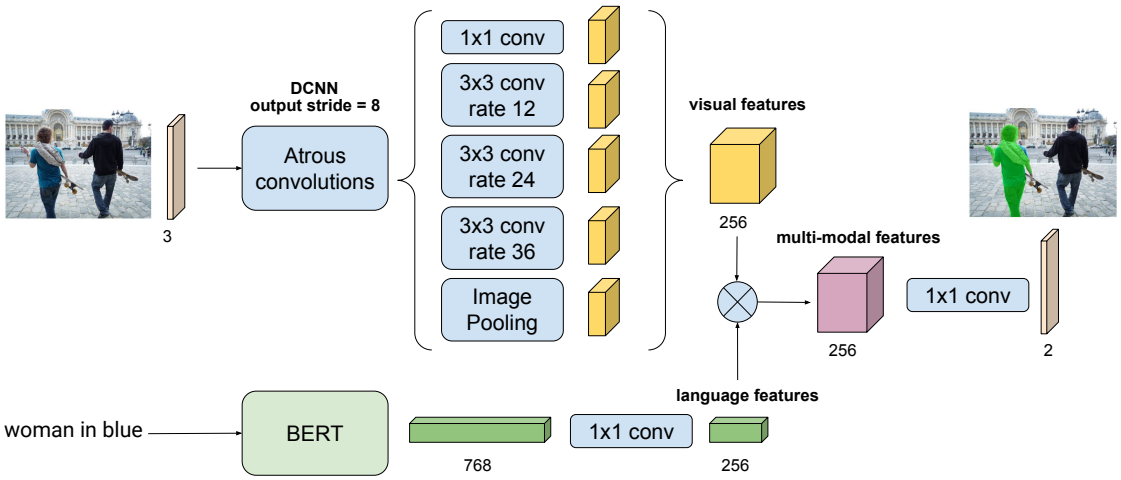


Figure 2.3: Architecture of the RefVOS model [Bell20].

resentations to downstream tasks were used: (i) *feature-based*, such as ELMo [Pete18] and (ii) *fine-tuning*, such as the Generative Pre-trained Transformer (OpenAI GPT) [Radf18].

The aforementioned approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations. This means that they look at a text sequence either from left to right or combined left-to-right and right-to-left while training. In contrast, BERT is applying a bidirectional training and its performance in several downstream tasks shows that a language model which is bidirectionally trained can have a deeper sense of language context and flow than a single-direction language model.

BERT makes use of Transformer [Vasw17], an attention mechanism that learns contextual relations between words (or sub-words) in a text. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once, therefore it is considered bidirectional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). In order to achieve that, the authors of BERT, use a "masked language model" (MLM) pre-training objective, inspired by the Cloze task [Tayl53]. The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM objective enables the representation to fuse both left and right context, which allows a bidirectional pre-training.

Sentences given as input to BERT are transformed to token sequences. The first token of every sequence is always a special classification token ([CLS]). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. Since in the task of referring image/video object segmentation the whole sentence is important for the identification of the referred object, RefVOS uses the learned embedding corresponding to the [CLS] token as its linguistic representation, which is subsequently combined with the visual features.

The output of BERT is a 768-dimensional vector for each token of the sequence. To obtain a multi-modal embedding, the encoded linguistic phrase is first converted to a 256-dimensional embedding through a linear projection, in order to match the number of extracted visual features from DeepLabv3, *i.e.* the feature maps. Then, the linguistic embedding is element-wise multiplied with the visual features at every pixel position, producing a multi-modal embedding. Finally, a 1×1 convolutional layer predicts two maps, one for the foreground class, *i.e.* the referent, and another for the background.

2.2.2 Relevant datasets

DAVIS-2017

The first dataset combining video object segmentation and referring expressions was DAVIS-2017 [Pont17], where the name DAVIS stands for "Densely Annotated Video Segmentation". The first

DAVIS dataset was presented in 2016 by Perazzi *et al.* [Pera16], being the first benchmark specifically designed for the task of video object segmentation, including 50 videos with one pixel-wise annotated object in each. In 2017 a new version of DAVIS was released by Pont-Tuset *et al.* [Pont17], which, besides having a bigger number of sequences (100 additional videos), was modified to include multiple annotated objects in its videos, in contrast with the previous version. Of course, for the task of referring video object segmentation, which is tackled in this work, DAVIS-2017 is a more suitable benchmark, since disambiguation between different classes of objects is one of the main challenges of this task.

DAVIS-2017 was annotated with referring expressions by Khoreva *et al.* [Khor18], who were the first to propose the replacement of the first frame mask supervision with a referring expression for the task of video object segmentation. They collected two different types of annotations from two annotators: (i) first frame annotations which are the ones produced by only looking at the first frame of the video and (ii) full video annotations which are produced after seeing the whole video sequence. The annotation procedure involved a non-interactive referential two-player game setup. A first annotator was asked to provide a language description of the object, which has a mask annotation by looking either at the first frame or at the full video (according to the type of annotation previously described). Then another annotator is given the first frame or full video and the corresponding description, and is asked to identify the referred object. If the second annotator correctly identifies the target object the expression is accepted, otherwise, it is corrected to remove ambiguity and to specify the object uniquely.

The augmented with referring expressions DAVIS-2017 contains 1,544 referring expressions for 386 unique objects appearing in 150 videos. The average length for the first frame and full video expressions is 5.5 and 6.3 words respectively. Although the videos of DAVIS-2017 consist of a large number of annotated frames (69.7 on average) in comparison to other relevant datasets, its validation set (which is used as a test set) is much smaller than the respective of other datasets, including only 30 videos. For this reason the experiments of the present work include also an evaluation on both the training and validation sets of DAVIS-2017 (90 videos in total), for the models which are not fine-tuned on this dataset.

A2D Sentences

Another dataset used in language-guided video object segmentation is A2D Sentences, created by Gavrilyuk *et al.* [Gavr18]. This dataset is based on the Actor-Action Dataset (A2D) [Xu15], which is a benchmark for action understanding consisting of 3,782 videos from YouTube. It includes seven annotated actor classes considered to perform a set of eight possible actions. A2D Sentences is the augmented version of A2D with natural language descriptions, stating what each actor is doing in each video.

The creators of the dataset, following the guidelines of RefCOCO dataset [Kaze14], ask the annotators for a discriminative referring expression of each actor instance if multiple objects are present in a video. A2D Sentences is finally composed of 6,656 sentences for 3,782 videos and 4,825 objects. Its sentences contain on average more words than the extended with referring exprssions DAVIS-2017 [Khor18] (7.3 versus 5.9). Since it is a dataset targeted for action description, its sentences emphasize on verbs having a total of 225 different verbs.

Refer-YouTube-VOS

The last video dataset augmented with referring expressions is Refer-YouTube-VOS, created by Seo *et al.* [Seo20] who collected crowd-sourced referring expression for YouTube-VOS [Xu18] using Amazon Mechanical Turk. YouTube-VOS is the largest existing benchmark for video object segmentation, including more than four thousand high-resolution videos collected from YouTube with a small duration of three to six seconds each. It includes pixel-level mask annotations for 94 different object categories at every five frames, while its videos have a frame rate of 30 frames per second.

In order to collect crowd-sourced referring expressions, the authors of Refer-YouTube-VOS firstly selected around 50 annotators after performing a validation test. Each annotator was given a pair of videos, the original video and the mask-overlaid one with the target object highlighted, and was asked to provide a discriminative sentence within 20 words that describes the target object accurately. Similar to Khoreva *et al.* [Khor18], two types of annotations were collected, one based on the first-frame and one on the full video. After the initial annotation, a verification and cleaning step was conducted for all annotations, and objects which could not be localized using just the produced language expression, were excluded from the dataset. In the end, Refer-YouTube-VOS consists of 27,899 expressions, referring to 7,451 objects in 3,975 videos, being the largest dataset with referring expressions in the video domain. Finally, Refer-YouTube-VOS has the largest average number of words per referring expression which is 7.5 for the first-frame annotations and 10.0 for the full-video ones.

Since YouTube-VOS, the basis of Refer-YouTube-VOS, and YouTube-VIS, the basis of the present work's proposed synthetic dataset, have a high overlap in their videos, the subset of Refer-YouTube-VOS that corresponds to YouTube-VIS has served as a benchmark for a direct comparison of the human-produced referring expressions with the respective synthetic ones proposed in the present Master thesis.

2.3 Object Detection

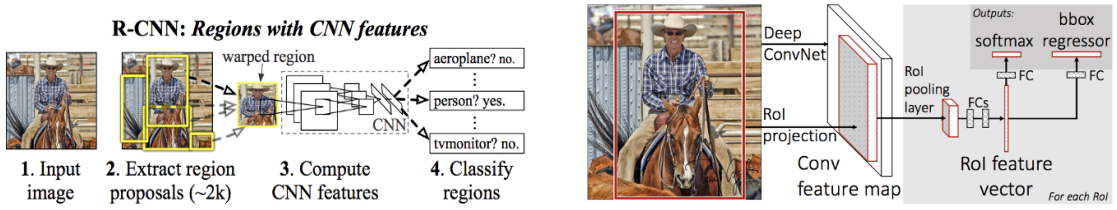
Object detection is the task of locating and classifying existing objects of a certain semantic class, as well as labeling them with rectangular bounding boxes which show the confidence of their existence. Being a classic computer vision problem, before the deep learning revolution in 2010s object detection has been approached with other machine learning-based methods. These methods first extract hand-crafted features like Haar [Viol01] or HOG [Dala05] features and SIFT keypoints [Lowe99] and then use machine learning techniques such as SVMs [Cort95] to do the classification.

However, after the recent advancements in deep learning, CNN-based methods have pushed the state-of-the-art in object detection as these techniques are able to detect objects in an end-to-end fashion without specifically defining features, outperforming classic computer vision methods in terms of detection accuracy. The frameworks of deep learning-based object detection methods can be mainly categorized into two types. The first one follows the traditional two-stage object detection pipeline, generating region proposals at first and then classifying each proposal into different object classes. The second considers object detection as a regression or classification problem, adopting a unified framework to acquire final object classes and locations in one step (single-stage detectors).

Two-Stage Detectors

Regarding two-stage detectors, the first stage is called a Region Proposal Network (RPN). A RPN takes an image (of any size) as input and outputs a set of rectangular object proposals, each with an objectness score, which measures the proposal's membership to a known set of object classes versus the background. Two-stage object detectors were introduced in the Selective Search work [Uijl13], while R-CNN [Girs14] was the first work to upgrade the second-stage classifier to a convolutional neural network yielding large gains in accuracy and introducing the deep learning era of object detection. The RPN of R-CNN extracts nearly 2000 region proposals, warps them into a square and feeds them to a convolutional neural network that produces a 4096-dimensional feature vector as output. Finally, a SVM takes as input these features acting as a classifier which decides on the presence of the object within that candidate region proposal. Besides predicting the presence of an object within the region proposal, the algorithm also predicts four values which are offset values to increase the precision of the bounding box. The pipeline of R-CNN is illustrated in Figure 2.4a.

Fast R-CNN [Girs15] instead of inputting 2000 region proposals to the CNN, uses the CNN to generate a feature map from the input image. From the convolutional feature map, the region proposals are identified and warped into squares. Then, a region of interest (RoI) pooling layer is used to reshape



(a) R-CNN (b) Fast R-CNN

Figure 2.4: Overview of R-CNN [Girs14] and Fast R-CNN [Girs15] two-stage frameworks for object detection.

them into a fixed size so that they can be fed into a fully connected layer. The RoI pooling operation significantly speeds up the process as the same input feature map can be used for a big number of region proposals of different size to finally get a list of corresponding feature maps with a fixed size. Finally, from the RoI feature vector, a softmax layer is used to predict the class of the proposed region and a softmax regression layer to predict the bounding box coordinates. An overview of the Fast R-CNN network is depicted in Figure 2.4b.

Both of the above algorithms (R-CNN and Fast R-CNN) use selective search [Uijl13] to find out the region proposals, which is a slow and time-consuming process affecting the performance of the network. To face this bottleneck and eliminate selective search, Faster R-CNN by Ren *et al.* [Ren15] employs a fully-convolutional network as a separate region proposal network (RPN) which has the ability to predict object bounds and scores at each position simultaneously. In this way the region proposal stage acts in a nearly cost-free way by sharing full-image convolutional features with the detection network. More specifically, for every point in the output feature map of the fully-convolutional network, the RPN has to learn whether an object is present in the input image at its corresponding location and estimate its size. This is done by placing a set of k "anchors" on the input image for each location in the output feature map. These anchors are rectangles that indicate possible objects in various sizes and aspect ratios at this location. As the network moves through each pixel in the feature map, it has to check whether these k corresponding anchors spanning the input image actually contain objects, and refine these anchors' coordinates to give bounding boxes as "object proposals" or regions of interest. Finally, these proposals are given to a Fast R-CNN [Girs15] object detector (Figure 2.4b) which predicts the class of the proposed region and also the bounding box coordinates. A high-level representation of the architecture of Faster R-CNN is illustrated in Figure 2.5.

The proposed method of the present Master thesis employs Faster R-CNN in order to detect attributes of the target objects which are used for the generation of synthetic referring expressions. As explained above and depicted in Figure 2.4b, the RoI feature vector of Fast R-CNN is guided to two sibling fully connected networks, one for the classification of the bounding box to the available object classes and a second for the prediction of the bounding box coordinates. These two branches are called "RoI heads" or just "heads". The proposed method for generating referring expressions uses Faster R-CNN extended with an attribute head by Tang *et al.* [Tang20] which is trained in order to detect a number of attributes for the detected objects like for example their color.

Subsequently to the R-CNN family of object detectors, other two-stage frameworks that have made an impact in object detection include R-FCN [Dai16] and FPN [Lin17]. R-FCN, while using a RPN similar to the one of Faster R-CNN [Ren15], modifies the classification network to a region-based fully convolutional detector where almost all computation is shared on the entire image, instead of applying a costly per-region subnetwork hundreds of times. The last convolutional layer of the detector produces position-sensitive score maps for each object class and then a position-sensitive RoI pooling layer is appended to aggregate the responses from these score maps and predict the class, while another convolutional layer is appended to obtain class-agnostic bounding boxes.

FPN (Feature Pyramid Network), from the other side, uses an architecture with a bottom-up path-

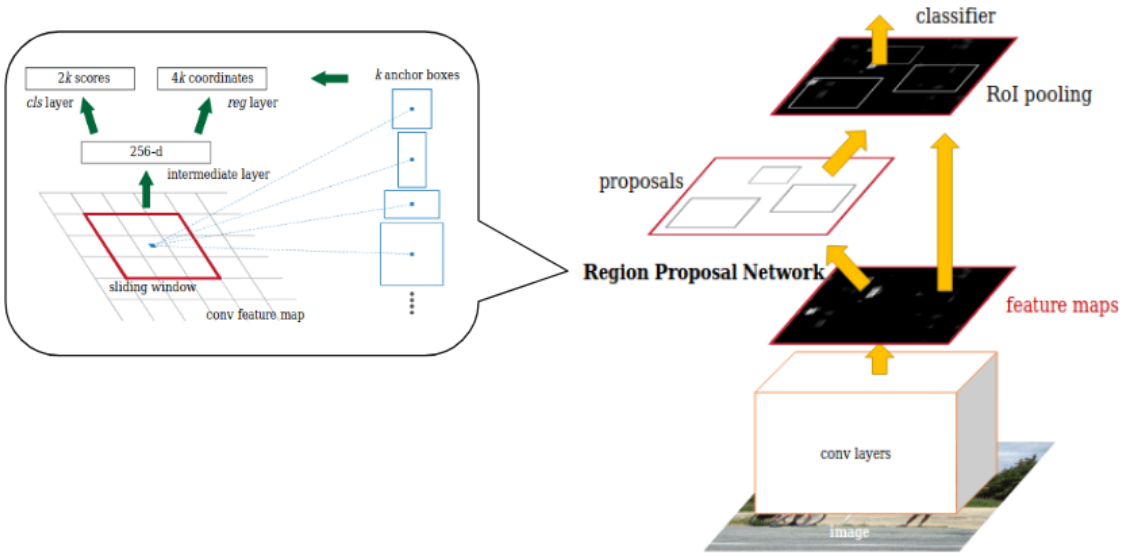


Figure 2.5: Architecture of Faster R-CNN [Ren15].

way, a top-down pathway and several lateral connections to combine low-resolution and semantically strong features with high-resolution and semantically weak features. While the bottom-up pathway is a basic backbone convolutional network, in order to build the top-down pathway, feature maps from higher network stages are upsampled at first and then enhanced with those of the same spatial size from the bottom-up pathway via lateral connections. Meanwhile, FPN is independent of the backbone CNN architecture and can be applied to different stages of object detection (*e.g.* region proposal generation) and to many other computer vision tasks (*e.g.* instance segmentation).

Single-Stage Detectors

The previously described two-stage frameworks are composed of several correlated steps, including region proposal generation, feature extraction, classification and bounding box regression, which are usually trained separately. Even in the end-to-end Faster R-CNN [Ren15], an alternating training is still required to obtain shared convolution parameters between the RPN and detection network. As a result, the time spent in handling different components becomes the bottleneck in real-time applications.

From the other side, single-stage frameworks based on global regression/classification, mapping directly from image pixels to bounding box coordinates and class probabilities, can significantly reduce training and testing time. The most successful and influential single-stage networks for object detection are YOLO [Redm16] and Single Shot MultiBox Detector (SSD) [Liu16]. YOLO (standing for "You only look once") makes use of the whole topmost feature map to predict confidences for both object categories and bounding boxes. The basic idea of YOLO is that it divides the input image into an $S \times S$ grid where each grid cell is responsible for predicting the object centered in it. Each grid cell predicts bounding boxes and their corresponding confidence scores. At the same time, regardless of the number of boxes, C conditional class probabilities are also predicted in each grid cell, corresponding to the known classes. The final bounding boxes are produced by combining the predicted bounding boxes and their confidences, for each grid cell, with the class probabilities. Although YOLO runs much faster than the state-of-the art two-stage object detectors, such as Faster R-CNN [Ren15], its detection accuracy is generally inferior as it has difficulty in dealing with small objects in groups or objects in unseen aspect ratios, which is caused by strong spatial constraints imposed on bounding box predictions and its relatively coarse features due to multiple downsampling operations.

Aiming at these problems, Liu *et al.*proposed a Single Shot MultiBox Detector (SSD) [Liu16], inspired by the anchors adopted in MultiBox [Erha14], the RPN of Faster R-CNN [Ren15] and multi-scale representations [Bell16]. Given a specific feature map, instead of fixed grids adopted in YOLO, the SSD takes advantage of a set of default anchor boxes with different aspect ratios and scales to discretize the output space of bounding boxes. To handle objects with various sizes, the network fuses predictions from multiple feature maps with different resolutions. By further leveraging techniques such as hard negative mining, data augmentation and a larger number of carefully chosen default anchors, SSD significantly outperforms Faster R-CNN in terms of accuracy on standard object detection benchmarks, while being three times faster.

2.4 Synthetic Data

Synthesizing training data has been explored in numerous applications in the fields of computer vision and natural language processing. The need of large amounts of data to train CNNs has encouraged the generation of synthetic datasets for solving tasks where real data cannot be easily collected. Flying Chairs [Doso15] and SURREAL [Varo17] are examples of synthetic datasets effectively used together with real data for the tasks of optical flow and human pose/shape estimation, respectively. Peng *et al.* [Peng15] augmented existing datasets for few-shot object detection by synthesizing images from freely available 3D CAD models of objects. While in the aforementioned works synthetic and real data are mixed, Saleh *et al.* [Sale18] proposed an effective way to use only synthetic data for semantic segmentation, by differentiating between foreground and background classes and using a detection-based approach. An enhanced Generative Adversarial Network (GAN) was used by Shrivastava *et al.* [Shri17] aiming to reduce the domain gap between real and synthetic images. The authors showed the effectiveness of their method in the tasks of gaze and hand pose estimation. In a more recent work, Khoreva *et al.* [Khor19] recommend a training strategy using fewer in-domain than large-scale out-of-domain data, by exploiting the provided annotation on the first frame of a video to synthesize realistic future video frames.

Synthetic linguistic data have also been used for training deep models on vision & language tasks. Fried *et al.* [Frie18] proposed a speaker-follower model to synthesize instructions for the task of vision-and-language-navigation. Silberer and Pinkal [Silb18] addressed the task of visual semantic role labeling and proved the effectiveness of training with synthetic data automatically created by applying a natural language processing model to image captions.

A highly related work to the present Master thesis is PhraseCut [Wu20]. Its authors address the task of language-guided image segmentation and create synthetic referring phrases for the images of

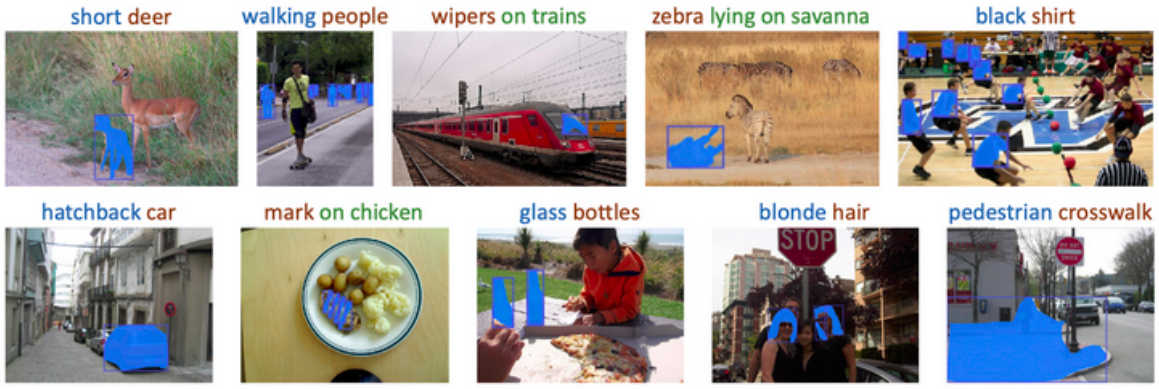


Figure 2.6: Examples from the PhraseCut dataset [Wu20]. Referring phrases are produced by combining object categories (brown text), attributes (blue text) and relationships (green text) with other objects.

Visual Genome dataset [Kris17] by combining the ground-truth (annotated by humans) object categories, attributes and relationships between them. Examples from the PhraseCut dataset are provided in Figure 2.6. Since the dataset targets the segmentation of image regions and not only objects, the generated expressions can refer to multiple objects, therefore they cannot be considered referring expressions, according to the definition given in Section 1.2. Another main difference between PhraseCut and the work of the present Master thesis is that the proposed method of the present work is dataset-independent and can be applied to any existing dataset that includes labeled object categories and bounding boxes.

Chapter 3

Proposed Method and Generated Dataset

3.1 YouTube-VIS Dataset

The dataset used for the generation of synthetic referring expressions is YouTube-VIS [Yang19], which is created on top of the large-scale video object segmentation dataset called YouTube-VOS [Xu18]. YouTube-VOS is the largest existing benchmark for video object segmentation, including more than four thousand high-resolution videos collected from YouTube with a small duration of 3-6 seconds each. Although YouTube-VOS contains pixel-level mask annotations for 94 different object categories, the reason that YouTube-VIS was preferred for creating synthetic referring expressions, is that the former is not exhaustively annotated, meaning that not all objects appearing in a video (belonging to those 94 categories) have a corresponding bounding box and segmentation mask annotation.

In contrast, YouTube-VIS, despite having a smaller category set of 40 common objects, it has the advantage that all instances belonging to those categories are labeled. In this way it serves as a very good data source for the task of generating synthetic referring expressions, as it is necessary to combine the information of all the present objects in a video frame in order to create valid referring expressions. YouTube-VIS totally consists of 2,883 videos with 4,883 unique objects belonging to 40 categories and approximately 131K object masks. However, since ground-truth annotations for all the frames are necessary to apply our proposed method, only the training set of YouTube-VIS can be used for this task which includes 2,238 videos with 3,374 annotated objects appearing in them.

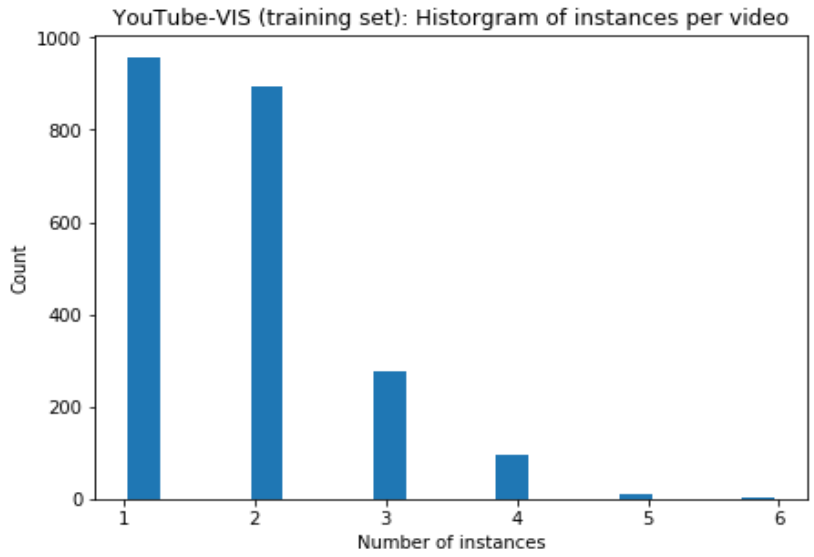


Figure 3.1: Histogram of object instances per video for the YouTube-VIS [Yang19] dataset.

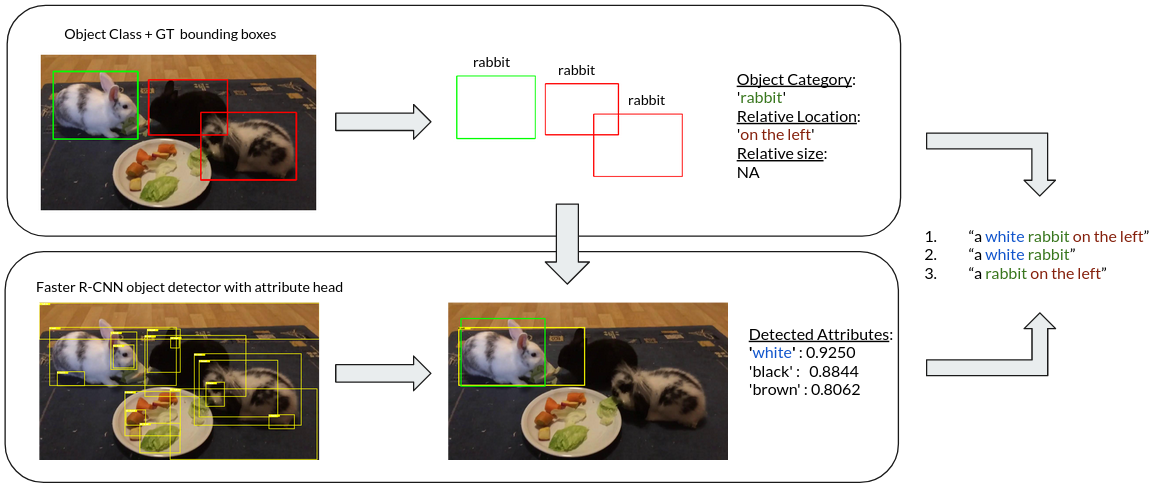


Figure 3.2: Overview of the proposed method for generating synthetic referring expressions. Top: Ground truth labels (object class + bounding boxes) are used to compute a target object's relative location and size. Bottom: A Faster R-CNN object detector with attribute head outputs attributes for the detected objects, which are filtered by the ground truth annotations. The combined cues create a set of referring expressions that uniquely identify the target object.

3.2 Proposed Method

The proposed method takes advantage of the ground-truth annotations of YouTube-VIS [Yang19] in order to generate synthetic referring expressions for the objects appearing in its videos. Specifically, the classes and bounding boxes of the target and other objects present in the frame are used in order to determine a set of cues which, when combined, are able to generate a referring expression that is close to a natural language expression. Also, Faster R-CNN [Ren15] (described in Section 2.3), enhanced with an attribute head by Tang *et al.* [Tang20], is employed for detecting attributes of the target object. The overview of the method and the cues used for generating synthetic referring expressions are illustrated in Figure 3.2. Four different cues are leveraged for generating a synthetic referring expression for a target object: (i) object class, (ii) relative size, (iii) relative location and (iv) attributes.

3.2.1 Object class

As already mentioned, in trivial cases where a single object of a known class is present, using the object's class is enough in order to generate a referring expression. However, the majority of cases involves multiple objects of the same class, thus other cues are necessary in order to disambiguate between instances. Relative size and location of a target object with respect to other objects of the same class are cues that can be easily computed using their bounding boxes.

3.2.2 Relative size

Relative size can be important in scenarios where multiple objects of the same class with similar characteristics are present. In the proposed method, in order to compute the relative size of a target object, two scenarios are considered:

1. If there is only one more object of the same class as the target, the areas of the bounding boxes of the two objects (target and other) are computed and compared. If the area A_t of the target object's bounding box compared to the area A_o of the other object's bounding box is twice as

big ($A_t >= 2A_o$) or small ($A_t <= 0.5A_o$), then a characterization of *"bigger"* or *"smaller"* is added to the synthetic referring expression respectively. Otherwise ($0.5A_o < A_t < 2A_o$), relative size is considered not applicable and is not included in the synthetic referring expression.

2. If there are $N >= 2$ other objects of the same class as the target, the area A_t of the target object's bounding box is compared to the areas A_o^i for $i = 1, ..., N$ of all the other objects' bounding boxes. Then, only if the area of the target object's bounding box is two times bigger or smaller from the areas of each of the other objects' bounding boxes ($A_t^i >= 2A_o^i$ for $i = 1, ..., N$ or $A_t^i <= 0.5A_o^i$ for $i = 1, ..., N$) a characterization of *"the biggest"* or *"the smallest"* is added to the synthetic referring expression respectively.

3.2.3 Relative location

In scenarios where two or three objects of the same class are present in a particular video frame, relative location between these objects can be used in order to disambiguate between them. If the bounding boxes of the objects are fully separable, or partially above a certain threshold, then it is assumed that relative location of the referent with respect to the other object(s) of the same class can be used in order to generate a non-ambiguous referring phrase. In this case, the steps for determining relative location are the following:

1. The pixel indices of the boundaries of the target and the other object's bounding boxes are considered in order to determine which axis (X or Y) is the most separative between them.

2. Then, three scenarios are considered:

   (i) Bounding boxes are fully separable on the determined axis: If X-axis is the most separative and the target object's bounding box pixel indices on the X-axis are smaller than those of the other object's, then relative location will be *"on the left"*. Otherwise, if the indices are bigger, it will be *"on the right"*. If Y-axis is the most separative and the target object's bounding box pixel indices on the Y-axis are smaller than those of the other object's, then relative location will be *"in the back"*, otherwise, *"in the front"*.

   (ii) Bounding boxes are partially separable on the determined axis: The degree of separation between the two bounding boxes is calculated by finding the maximum non-overlapping distance between the two bounding boxes. If this value is above a fixed threshold of 50 pixels, then relative location is applicable and one of the four options mentioned above is selected, according to the determined axis and the location of the boundaries. If the maximum non-overlapping distance is smaller than 50 pixels, relative location is not applicable.

   (iii) Bounding boxes are not separable: This implies that one bounding box is enclosed inside the other. Relative location is not applicable in this case.

3. If there are two other objects of the same class, besides the referent, steps 1 & 2 are computed between the referent and each of the two other objects and the results are combined. In such a case, if the referent is located, for example, on the right of the first other object and on the left of the second one, then its relative location will be *"in the middle"*. In a similar way more combinations of the 4 basic relative locations mentioned in step 2 can occur, *e.g. "in the front left", "in the back right"*, *etc*.

While the choice of *"left"* and *"right"* for the X-axis is trivial, *"back"* and *"front"* were selected for the Y-axis as they were found to be the most frequently used words for determining relative location in the Y-axis in referring expressions of DAVIS-2017 [Khor18] and A2D Sentences [Gavr18].

3.2.4 Attributes

Attributes like the color of an object have been proved to be important for the task of referring image/video object segmentation [Bell20]. In order to detect attributes for a target object, the proposed method employs Faster R-CNN [Ren15] object detector enhanced with an attribute head by Tang *et al.* [Tang20]. Faster R-CNN is pre-trained on Visual Genome [Kris17] with its attribute head enabled, so that the model is able to predict attributes, like color, for the detected objects. Then, the pre-trained model is run on YouTube-VIS [Yang19] to obtain, for each frame of a video, a set of detected objects (with their bounding box coordinates) and their detected attributes. For each detected bounding box from Faster R-CNN, its overlap with the referent's ground truth bounding box is computed using their Intersection-over-Union (IoU) which corresponds to the intersection area (in pixels) of the bounding boxes divided by their union area (a detailed description of IoU is provided in Section 4.2). The bounding box with the highest overlap is considered as the prediction which corresponds to the target object, with the condition that IoU is over 50%. The procedure is visually explained in Figure 3.2.

The attributes predicted for the selected bounding box are filtered to color-like and non color-like and the ones with the highest prediction score, if above 85%, are selected for the two subsets. For color-like attributes, if the scores of the first two colors are very close, *i.e.* their score difference is lower than 2%, both colors are used in the referring expression, since in many cases more than one color is necessary to describe an object (*e.g. "a yellow and green parrot"*). For non color-like attributes only the one with the highest score is selected. Non color-like attributes can be both adjectives (*e.g. "large", "spotted"*) or verbs (*e.g. "walking", "surfing"*). The model is able to detect a total of 201 attributes. An attribute is added to the referring expression of a target object only if no other objects belonging to the same class have this attribute, so that the final expression satisfies the definition of a referring expression which is to uniquely describe a target object.

3.2.5 Synthetic Referring Expressions

Finally, the aforementioned cues are combined in a natural order and a proper article is added to the sentence, ending up with a complete synthetic referring expression. There might be cases where none of the above cues are applicable for a target object and the generated synthetic language expression may be ambiguous, although in the vast majority of cases the synthetic language expressions uniquely identify the target object.

Since a video consists of a certain number of frames, and an object may change its location or

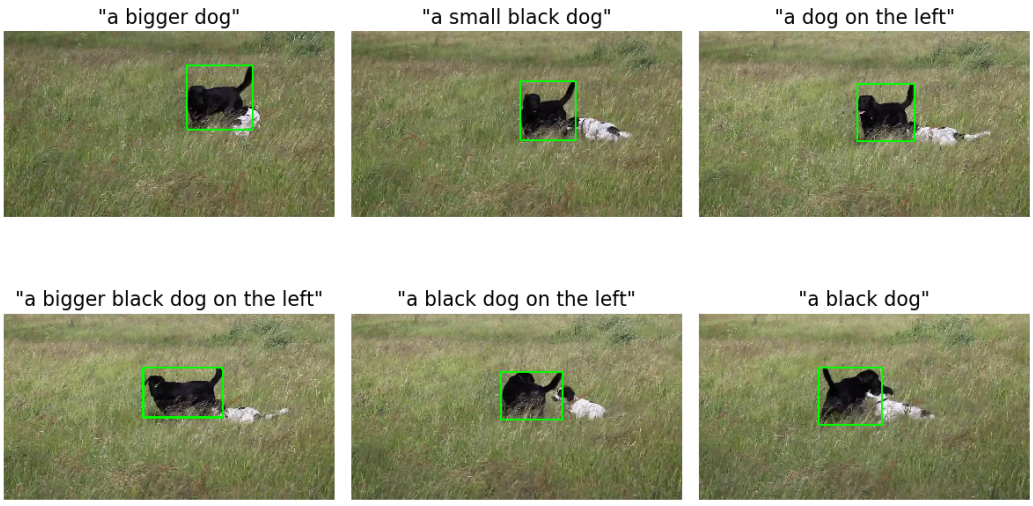


Figure 3.3: Example of synthetic referring expressions automatically generated with the proposed method. Multiple referring expressions can be created for the same video or even for the same frame.

appearance throughout the video, one or more synthetic referring expressions are generated for each frame of the video. In this way, a network can be trained with different referring expressions for the same video or for the same frame, increasing its ability to generalize to other data. An example of synthetic referring expressions generated with the proposed method for different frames of a video is illustrated in Figure 3.3.

3.3 SynthRef-YouTube-VIS Dataset

Only a few datasets are available for the task of referring video object segmentation. As mentioned in Section 2.2, Khoreva *et al.* [Khor18] and Gavrilyuk *et al.* [Gavr18] have augmented the DAVIS-2017 [Pont17] and A2D [Xu15], J-HMDB [Jhua13] datasets respectively. However, the limited number of videos of the former and the restricted set of object categories of the latter make them unsuitable for effectively pre-training a deep neural network for the task of referring video object segmentation. In a concurrent work to the present Master thesis, Seo *et al.* [Seo20] presented Refer-YouTube-VOS, which is the first large-scale dataset created for the task of referring video object segmentation. They employed Amazon Mechanical Turk to collect referring expressions for YouTube-VOS [Xu18], which consists of a large number of videos and object categories.

One can understand that annotating a dataset such as YouTube-VOS, which includes 4,519 videos and nearly 7,500 objects, involves a big annotation cost in terms of money and/or time. On the contrary, the proposed dataset of synthetic referring expressions, which we call SynthRef-YouTube-VIS, is created without any additional human annotation cost. SynthRef-YouTube-VIS is based on YouTube-VIS [Yang19], a subset of YouTube-VOS [Xu18], originally used for the task of video instance segmentation. The method used for generating SynthRef-YouTube-VIS is described in detail in Section 3.2, while a qualitative comparison of the synthetic referring expression of SynthRef-YouTube-VIS and the human-produced ones of Refer-YouTube-VOS for the same videos is illustrated in Figure 3.4.

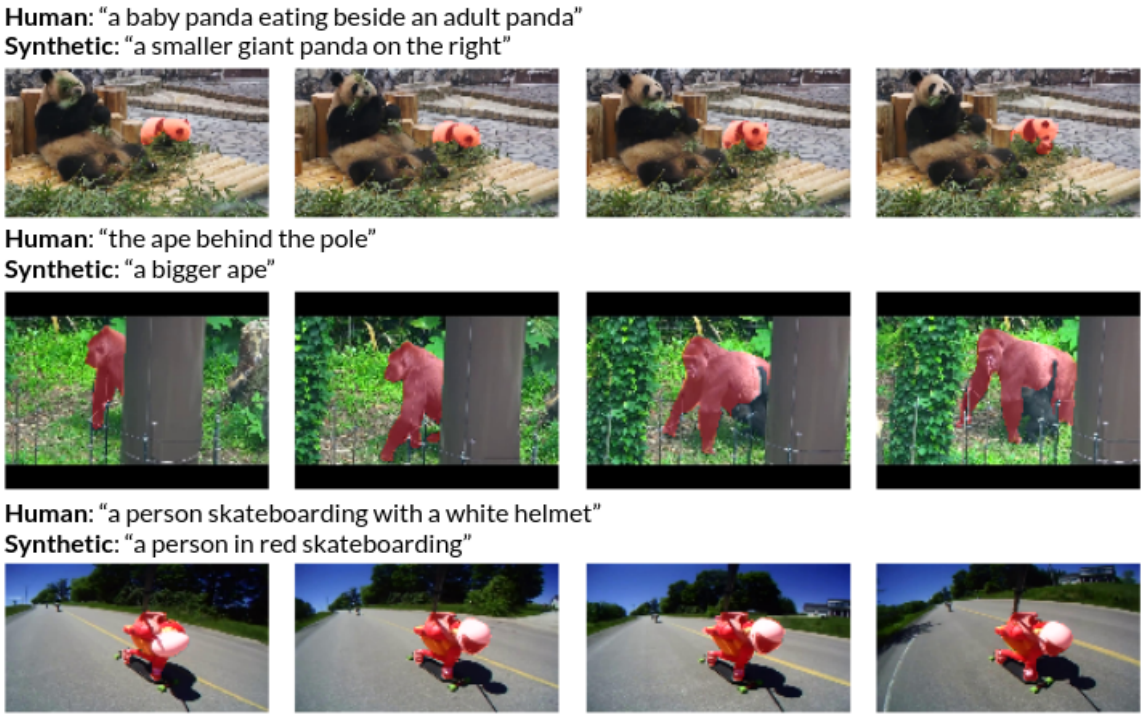


Figure 3.4: Comparison of human-produced referring expressions of Refer-YouTube-VOS [Seo20] with synthetic ones generated with the proposed method.

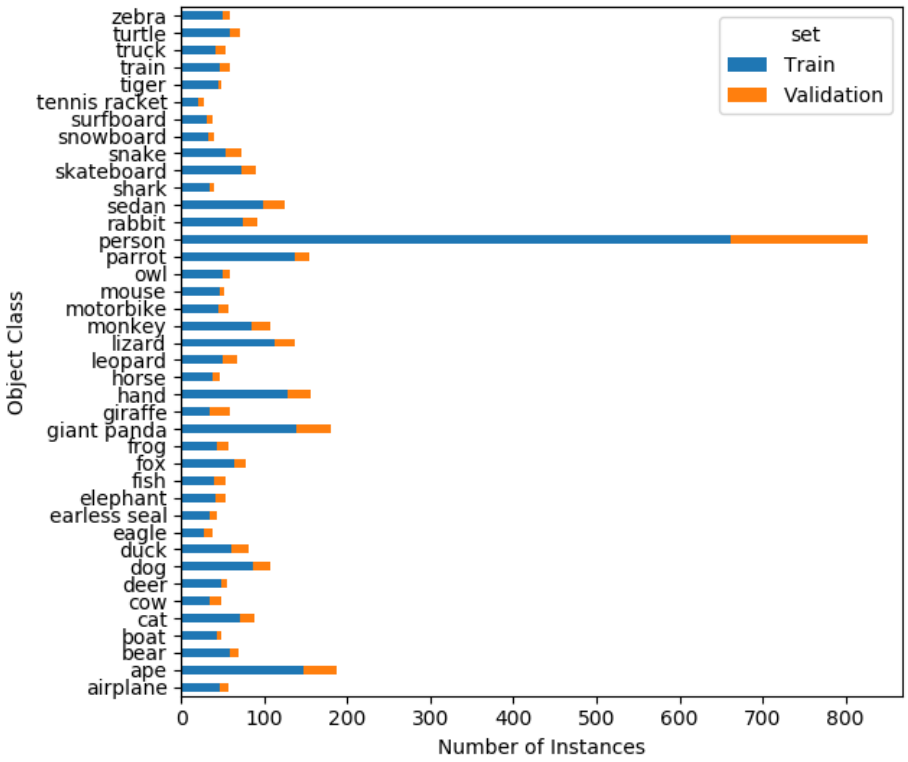


Figure 3.5: Histogram of object instances by each class in SynthRef-YouTube-VIS train and validation sets.

The official training set of YouTube-VIS is used for the creation of SynthRef-YouTube-VIS, since ground-truth annotations for all the frames are necessary in order to apply our proposed method. In this way, SynthRef-YouTube-VIS consists of 2,238 videos with 3,374 annotated objects appearing in them. The dataset is further split in a train and test set for the experiments having 1791 training and 447 testing videos. As seen in Figure 3.5, all classes appear both in the training and the test set, in contrast to DAVIS-2017 [Pont17] where there exist object classes in the validation set that are never seen during training (see Figure 4.2 of Section 4.3).

A statistical analysis and comparison between the proposed synthetic dataset and other referring video object segmentation datasets, which have been described in detail in Section 2.2.2, is illustrated in Table 3.1. J-HMDB [Jhua13] is not included in the analysis since its videos only contain one annotated object and thus is not suited for multi-instance segmentation.

Dataset	**Videos**	**Objects**	**Classes**	**RE Type**	**REs**	**REs/Object**
A2D Sentences [Gavr18]	3,782	4,825	8	*Human*	6,656	1.4
DAVIS-2017 [Khor18]	150	386	78	*Human*	1,544	4.0
Refer-YouTube-VOS [Seo20]	3,975	7,451	94	*Human*	27,899	3.7
SynthRef-YouTube-VIS (Ours)	2,238	3,774	40	*Synthetic*	15,798	**4.2**

Table 3.1: Statistics of the proposed dataset and comparison to existing ones. The last column indicates the average number of unique referring expressions (REs) per object.

As depicted in Table 3.1, the proposed synthetic dataset includes a total of 15,798 unique referring expressions for all 3,774 objects of YouTube-VIS training set. This number is quite higher than the respective number of A2D Sentences [Gavr18] and DAVIS-2017 [Khor18]. Although it is smaller than

Refer-YouTube-VOS [Seo20] in terms of videos and total number of expressions, SynthRef-YouTube-VIS still has the highest average number of unique referring expressions per annotated object, which is 4.2. Finally, the average number of words of the synthetic referring expressions of SynthRef-YouTube-VIS is 4.4, which is smaller than the respective of the other datasets. This is reasonable as the goal of the proposed method is to generate simple and efficient synthetic referring expressions, using only the previously described cues (object class, relative size/location and attributes).

Chapter 4

Experimental Results

4.1 Training Details

The experiments of the present work intend to assess the benefits of training a model for the task of referring video object segmentation using the synthetic referring expressions generated with the proposed method. The model used in the experiments is RefVOS [Bell20], which is described in detail in Section 2.2. Two types of experiments are conducted:

1. The first experiment consists of using the generated synthetic dataset, SynthRef-YouTube-VIS, as an extra dataset for training a model which is already pre-trained with real (*i.e.* human-produced) referring expressions and evaluating its performance on DAVIS-2017 [Khor18] and A2D Sentences [Gavr18].

2. In the second experiment, the model is trained, on the one hand, using only the proposed synthetic data and, on the other hand, using only real data. Both models are evaluated on the real referring expressions of Refer-YouTube-VOS [Seo20] in order to compare the performance of training on human versus synthetic referring expressions, on the same dataset.

Finally, an ablation study of different settings while pre-training with the synthetic referring expressions is also presented as well as an analysis of the impact of the information included in the referring expressions on the segmentation accuracy.

4.1.1 Pre-training

Pre-training a deep neural network on a large dataset before fine-tuning it on a smaller one is a common technique used in deep learning. As explained in the previous chapters, the present work assesses how human and synthetic referring expressions can be used together or interchangeably for pre-training a model for the task of referring video object segmentation. The two datasets used in the experiments for pre-training are RefCOCO [Kaze14] and SynthRef-YouTube-VIS, which is the synthetic dataset generated with the proposed method.

For pre-training the RefVOS model [Bell20], a batch size of eight video frames is used, which are resized and then cropped/padded to a final resolution of 480x480. The large crop size is necessary for the visual encoder (DeepLabv3 [Chen17b]), so that atrous convolutions (see Section 2.2 for details) with large rates are effective. Otherwise, the filter weights with a large atrous rate are mostly applied to the padded zero region of the image or frame. The loss function employed is the binary cross-entropy loss since the model predicts two classes, one for the foreground which corresponds to the target object and one for the background. The optimizer employed is stochastic gradient descent (SGD) with a momentum of 0.9.

The learning rate values and schedule depend on the dataset and the training step, *i.e.* if the model is already trained on other data or not. When pre-training from scratch on RefCOCO or SynthRef-YouTube-VIS, an initial learning rate is set to 0.01 and is decreased by 4×10^{-4} at every epoch for a total of 24 epochs. For training on SynthRef-YouTube-VIS after a first pre-training on RefCOCO, a smaller learning rate of 10^{-4} is used, which is linearly decreased by 4×10^{-6} after every epoch for 20 epochs.

4.1.2 Fine-tuning on the Evaluation Datasets

After the pre-training phase and before evaluating the model on a target dataset, it is a common practice to also train the model on the target dataset, a process which is called fine-tuning. The experiments conducted assess the model's performance both when fine-tuning or not on the target dataset before the evaluation. The batch size, optimizer and loss function are the same as the ones used in the pre-training phase of the model and only the learning rate policy is adjusted to the target dataset.

For the evaluation of the proposed method and dataset three benchmarks on video object segmentation are used, which have been further extended with referring expressions from previous works. Two of these benchmarks are used for the first experiment where RefVOS (the model used) is pre-trained either on RefCOCO [Kaze14] or SynthRef-YouTube-VIS or both of them. The first dataset is DAVIS-2017 [Khor18] and the second is A2D Sentences [Gavr18]. For fine-tuning on DAVIS-2017 the learning rate starts from 10^{-5} and is decreased to 10^{-6} after 10 epochs, training for a total of 15 epochs. For A2D Sentences, the learning rate is set to 10^{-4} and is linearly decreased by 4×10^{-6} after every epoch, for 20 epochs in total. The third dataset is Refer-YouTube-VOS [Seo20] and it is used for the second experiment as described above. Since in this experiment the model is not pre-trained on RefCOCO, the same learning rate policy as when pre-training with SynthRef-YouTube-VIS is used, which is, starting the learning rate from 0.01 and decreasing it by 4×10^{-4} at every epoch for a total of 24 epochs.

4.2 Quantitative Evaluation Metrics

In the task of object segmentation, given a ground truth mask $\mathcal{G}$ and a predicted segmentation mask $\mathcal{M}$, the typical evaluation process includes two measures, as proposed by Perazzi *et al.* [Pera16]:

1. Region Similarity $\mathcal{J}$: The similarity of the ground truth and predicted segmentation regions is measured using the Jaccard Index $\mathcal{J}$ defined as the *Intersection-over-Union (IoU)* of the two regions i.e.:
$$\mathcal{J} = \frac{|\mathcal{M} \cap \mathcal{G}|}{|\mathcal{M} \cup \mathcal{G}|}$$
2. Contour accuracy $\mathcal{F}$: The predicted segmentation mask $\mathcal{M}$ can be interpreted as a set of closed contours $c(\mathcal{M})$ delimiting the spatial extent of the mask. Then, the contour-based precision and P_c and recall R_c between the contour points of $c(\mathcal{M})$ and $c(\mathcal{G})$ can be computed using a bipartite graph matching, which is approximated via morphology operators for efficiency [Pera16]. The final accuracy is the typical F-measure (or F_1 score) defined as:
$$\mathcal{F} = \frac{2P_cR_c}{P_c + R_c}$$

Based on the above measures, the following metrics are being used in the experiments for the evaluation of the proposed method and the comparison to existing approaches:

- *Precision@X* : Given a threshold *X* in the range [0.5,0.9], a predicted mask for an object is counted as true positive if its $\mathcal{J}$ is larger than *X*, and as false positive otherwise. Then, Precision@X is computed as the ratio between the number of true positives and the total number of instances.
- *Overall* $\mathcal{J}(IoU)$: Total intersection area of all objects divided by the total union area.
- *Mean* $\mathcal{J}$ (IoU): Average of the $\mathcal{J}$ measure (IoU) of all objects so that large and small regions are treated equally.
- $\mathcal{J}\&\mathcal{F}$: The average of the mean region based similarity (Mean $\mathcal{J}$ and the mean contour accuracy (Mean $\mathcal{F}$).

The evaluation metrics in each experiment are selected according to the target dataset, so that a comparison with previous works is possible.

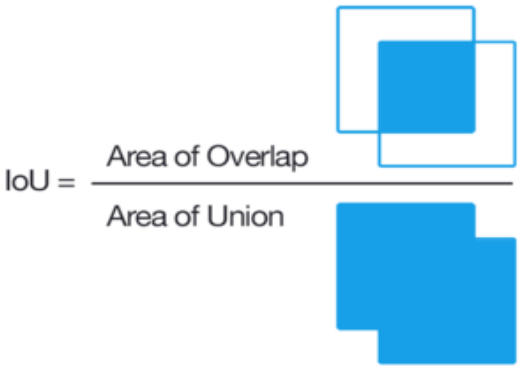


Figure 4.1: Visual explanation of the Intersection-over-Union (IoU) or Jaccard Index (J)

4.3 Quantitative Results

Quantitative results, which are organized according to the target dataset, are presented below. The DAVIS-2017 [Khor18] and A2D Sentences [Gavr18] datasets correspond to the first experiment whereas Refer-YouTube-VOS [Seo20] corresponds to the second.

DAVIS-2017 Validation

Results obtained in DAVIS-2017 validation set are compared with previous works from Khoreva *et al.* [Khor18], Seo *et al.* [Seo20] and Bellver *et al.* [Bell20] in Table 4.1. Previous works follow a standard approach, that is, pre-training a model on RefCOCO [Kaze14] and then fine-tuning on DAVIS-2017. Performance is also assessed when the model is not fine-tuned on the target dataset but only pre-trained either with human or both human and synthetic referring expressions.

By adding the synthetic referring expressions of SynthRef-YouTube-VIS in the pre-training phase and without fine-tuning on DAVIS-2017, a significant gain of 4% is observed from the respective model of [Bell20] (40.8) which is pretrained only on RefCOCO. The obtained $J\&F$ of 44.8 also outperforms the best models provided by Khoreva *et al.* [Khor18] (39.3) and Seo *et al.* [Seo20] when their model is pretrained on RefCOCO (44.1). It should be underlined that both previous models have been fine-tuned on DAVIS-2017 in contrast to the proposed one. When the proposed pre-trained model is also fine-tuned on DAVIS-2017 performance slightly increases from 44.8 to 45.3.

Model	Pretrain RefCOCO	+Pretrain SynthRef-YouTube-VIS	+Ft DAVIS	J&F
RefVOS [Bell20]	✓			40.8
RefVOS (Ours)	✓	✓		**44.8**
Khoreva *et al.* [Khor18]	✓		✓	39.3
URVOS [Seo20]	✓		✓	44.1
RefVOS [Bell20]	✓		✓	45.1
RefVOS (Ours)	✓	✓	✓	**45.3**

Table 4.1: Quantitative results when pre-training with our synthetic referring expressions and evaluating on DAVIS-2017 validation set.

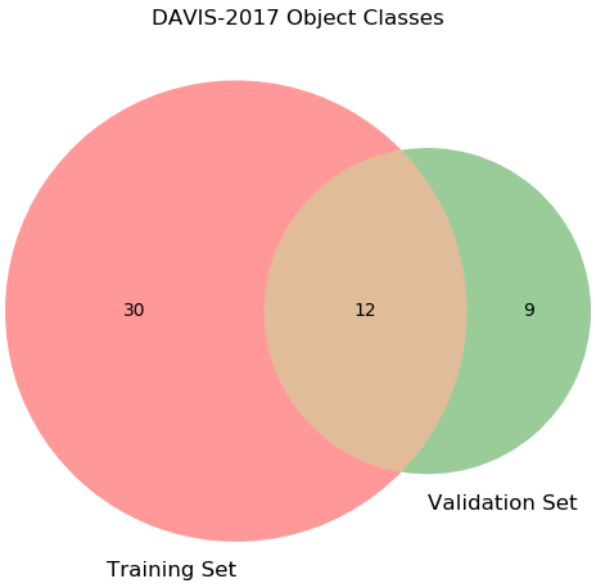


Figure 4.2: Venn diagram of the object classes in DAVIS-2017 training and validation sets.

DAVIS-2017 Training & Validation

Since the validation set of DAVIS-2017 consists of only 30 videos with 21 object classes, an evaluation on the 120 videos of the whole DAVIS-2017 (training plus validation sets) was also performed, for the models which were not fine-tuned on this dataset. As illustrated in the Venn diagram of Figure 4.2, DAVIS-2017 training set has 30 object classes that do not appear in the validation set. By evaluating on both training and validation splits of DAVIS-2017, a total of 51 object classes are included.

The results of this experiment are summarized in Table 4.2 confirming that pre-training with the generated synthetic dataset helps the model generalize better in a new set of data, as $J\&F$ increases 5 points, from 33.6 to 38.6. The performance when the model is pre-trained only on the synthetic referring expressions of SynthRef-YouTube-VIS is also assessed in the second row of Table 4.2. It is observed that the performance (27.0) is lower than when pretraining on RefCOCO (33.6), *i.e.* on real data, but it is still comparable, if one takes into account the annotation cost of the two datasets, which in this case (*i.e.* using only generated synthetic referring expressions) is equal to zero.

Pretraining	J&F
RefCOCO	33.6
SynthRef-YouTube-VIS	27.0
RefCOCO+SynthRef-YouTube-VIS	**38.6**

Table 4.2: Results on the training + validation sets of DAVIS-2017, without fine-tuning.

A2D Sentences

The same experiment is also conducted using the A2D Sentences dataset [Gavr18]. At first, the same model which was tested on DAVIS-2017, that is only pre-trained on RefCOCO [Kaze14] and SynthRef-YouTube-VIS, is evaluated on the A2D Sentences test set, without fine-tuning. Then, in a second setup, this model is also fine-tuned on A2D Sentences training set. Precision at several thresholds and the Overall and Mean IoU (J) are reported in order to be able to compare the obtained results with previous works.

Results are reported in Table 4.3. The first two rows represent the performance without fine-tuning on the target dataset (A2D Sentences). In this case it is observed that a second pre-training of the model using the proposed synthetic dataset increases the Mean IoU by 12 points (from 25.6 to 37.6), the Overall IoU by 8 points (from 41.4 to 49.4) as well as the Precision for all thresholds. This result confirms that the synthetic data generated with the proposed method help the model to generalize in a new dataset. This can be very important in a scenario where training data is not available for the target dataset due to money or time constraints, as an extra pre-training with the proposed synthetic dataset could give a significant increase to the segmentation accuracy.

The last two rows of Table 4.3 show the results obtained when further fine-tuning on the target dataset. It is noticed that the performance is increased when fine-tuning on the target dataset, both when synthetic data are used for pre-training or not (a result obtained by Bellver *et al.* [Bell20]). However, in contrast with DAVIS-2017 (Table 4.1), on A2D Sentences the synthetic data do not increase the performance when fine-tuning on the target dataset. This can be explained from the nature of A2D Sentences dataset and the type of its referring expressions. As already mentioned in Subsection 2.2.2, this dataset was created with the purpose of action description/recognition rather than object identification, thus its expressions are quite different than the respective of RefCOCO and SynthRef-YouTube-VIS, including mostly verbs and less attributes.

Training	Prec @0.5	Prec @0.6	Prec @0.7	Prec @0.8	Prec @0.9	Overall J	Mean J
RefCOCO [Bell20]	27.9	24.1	19.7	12.6	3.4	41.4	25.6
RefCOCO + SynthRef-YouTube-VIS	**42.8**	**36.0**	**27.0**	**15.8**	**3.5**	**49.4**	**37.6**
RefCOCO + ft. A2D [Bell20]	**57.8**	**53.1**	**45.6**	**31.0**	**9.3**	**67.2**	**49.7**
RefCOCO + SynthRef-YouTube-VIS + ft. A2D	54.0	47.8	37.9	22.9	5.0	64.1	45.4

Table 4.3: Results on A2D Sentences dataset confirm the advantage of pre-training with synthetic data when fine-tuning on the target dataset is not applicable.

Refer-YouTube-VOS

The second experiment focuses on comparing the generated synthetic referring expressions against the human-produced ones for the same videos. This comparison is achieved by using Refer-YouTube-VOS [Seo20] whose videos overlap with the respective ones from the proposed synthetic dataset, namely SynthRef-YouTube-VIS.

By using the subset of Refer-YouTube-VOS that corresponds to SynthRef-YouTube-VIS, two different models are trained: one model is trained using the human-produced referring expressions of Refer-YouTube-VOS, whereas a second model is trained using only the generated synthetic expressions of SynthRef-YouTube-VIS. The evaluation is done on the test split of SynthRef-YouTube-VIS but using the human-produced expressions of Refer-YouTube-VOS for both models in order to achieve a fair comparison. Since both human and synthetic referring expressions are available for the same videos, this result can be a measure of the domain gap between real and synthetic data for training.

Referring Expressions	Prec@ 0.5	Prec@ 0.6	Prec@ 0.7	Prec@ 0.8	Prec@ 0.9	Overall IoU	Mean IoU
Synthetic	32.27	24.05	16.30	8.48	1.82	40.12	35.02
Human	38.61	31.69	24.54	16.71	6.87	41.73	39.46

Table 4.4: Comparison of the performance on the subset of Refer-YouTube-VOS corresponding to SynthRef-YouTube-VIS, when training with synthetic and human referring expressions.

Results from this experiment are reported in Table 4.4. The results indicate that, even though the model trained on human referring expressions outperforms the model trained on synthetic ones, the drop in segmentation accuracy is not that big to prevent the use of the proposed synthetic data for training. On the contrary, the obtained numbers show that synthetic referring expressions generated with the proposed method can be used interchangeably with human ones when the latter are hard to acquire because of time and/or money constraints.

(a) Pretrained on RefCOCO (b) Pretrained on RefCOCO+SynthRef-YouTube-VIS

Figure 4.3: Qualitative results on DAVIS-2017. Subfigure 4.3a (left) shows results when the model is pre-trained only on RefCOCO, while Subfigure 4.3b (right) when it is also trained on the proposed synthetic dataset.

(a) Pretrained on RefCOCO (b) Pretrained on RefCOCO+SynthRef-YouTube-VIS

Figure 4.4: Qualitative results on A2D Sentences. The model in the left subfigure is pre-trained only on RefCOCO, while the model in the right subfigure is also trained with the generated synthetic referring expressions.

4.4 Qualitative Results

Qualitative results from DAVIS-2017 validation set [Khor18] are illustrated in Figure 4.3. These images are outputs of the experiment which corresponds to the first two rows of Table 4.1. The improvement of the segmentation masks for the referred objects is significant as is depicted in Subfigure 4.3b where the model is also pre-trained using the synthetic referring expressions. This is reflected both in the ability of the model to identify the referred instance (first and last row) but also in correctly segmenting the contour of the target object (second row).

Similar results for the A2D Sentences [Gavr18] dataset are depicted in Figure 4.4. The illustrated video frames correspond to the experiments reported on the first two rows of Table 4.3, where the model is either pre-trained only on RefCOCO [Kaze14], or it is also pre-trained with the proposed synthetic dataset. It can be easily observed that the segmentation results of the latter setup (Subfigure 4.4b) are superior compared to the the former.

4.5 Ablation Study

Synthetic Referring Expressions Analysis

In order to evaluate the effect of the information included in the synthetic referring expressions, experiments with different amount of information were conducted, starting from a baseline where the synthetic referring expressions consist of just the object class *e.g.* "a dog". Then in the second experiment, relative size and location are added and in the third and last experiment attributes are included too. The model used in these experiments is first pre-trained on RefCOCO, then trained with the synthetic referring expressions of SynthRef-YouTube-VIS, using different amount of information in each experiment, as explained above, and it is finally fine-tuned on the training set of DAVIS-2017 and evaluated on the validation set.

Results in Table 4.5 indicate that performance gradually increases with the amount of information provided in the synthetic referring expressions. This is explained because of the fact that in cases where multiple objects of the same class are present in a video, bigger amount of information is necessary in order to unambiguously identify a specific object. It is also remarkable that the final segmentation accuracy in DAVIS-2017 is high, even when only the object class is used as referring expression, during pre-training with SynthRef-YouTube-VIS. This happens for two reasons. The first is that the model is already pre-trained on RefCOCO and the second is that DAVIS-2017 validation set includes several videos where only one object instance from each class appears.

Referring Expression Information	J&F
Obj. Class	42.0
+ Relative Size + Relative Location	43.5
+ Attributes	45.3

Table 4.5: Effect of the information included in the synthetic referring expressions on the final performance on DAVIS-2017 validation set.

Freezing the language branch

A common approach when fine-tuning a model on a target dataset after firstly pre-training on another one, is to freeze some of the layers in order to avoid overfitting to the target data. This is also done for reducing the amount of time and memory a model needs to train, since by freezing some of the layers, less parameters need to be calculated.

In the present work, a freezing of the language encoder layers (*i.e.* BERT model [Devl19]) is assessed with the hypothesis that this way the model could avoid overfitting to the synthetic referring

expressions when it is already pre-trained on RefCOCO [Kaze14] (*i.e.* human-produced referring expressions). More specifically, the same configurations corresponding to the first experiment (as explained in Section 4.1) were repeated with freezing the language branch and fine-tuning only the visual one as well as the final layers after the multi-modal embedding is obtained.

However, results on DAVIS-2017 [Khor18] and A2D Sentences [Gavr18] datasets have shown that the effect on the segmentation accuracy is negligible. The results are summarized in Table 4.6. In DAVIS-2017, fine-tuning BERT while pre-training with the proposed synthetic data yields slightly better results, both when fine-tuning (the whole model) on the target dataset or not. On the contrary, in A2D Sentences the segmentation accuracy is slightly better when the language encoder layers are frozen during the pre-training using the synthetic referring expressions, regardless of whether the model is fine-tuned on A2D Sentences or not. The different behaviour than DAVIS-2017, as previously explained, can be justified from the fact that the proposed synthetic referring expressions are more similar to the ones of DAVIS-2017 than those of A2D Sentences, whose phrases intend to describe actions, containing a lot of verbs and less attributes. Nevertheless, similarly to DAVIS-2017, freezing or not freezing the language encoder when pre-training with synthetic expressions brings minor changes to the final segmentation accuracy (nearly 0.5%).

	DAVIS-2017 Val		DAVIS-2017 Train+Val	A2D Sentences	
	No Ft.	*Ft.*	*No Ft.*	*No Ft.*	*Ft.*
BERT frozen	44.7	45.0	38.2	**38.1**	**46.0**
BERT fine-tuned	**44.8**	**45.3**	**38.6**	37.6	45.4

Table 4.6: Analysis of the performance when freezing the language branch while pre-training on the proposed synthetic dataset. Results are split by target dataset and whether the model is fine-tuned on it (*Ft.*) or not (*No ft.*). Note that when fine-tuning on the target dataset the language branch is also fine-tuned. The reported metric for DAVIS-2017 is the J&F whereas for A2D Sentences it is the Mean J.

Chapter 5

Conclusions and Future Directions

This Master thesis proposes a simple yet effective method for automatically generating synthetic referring expressions for an image or video frame and creates the first large-scale dataset with synthetic referring expressions based on YouTube-VIS [Yang19], a dataset for video instance segmentation. Additionally, the synthetic dataset is evaluated by using it in the pre-training of a deep neural network for the task of referring video object segmentation. From the experiments presented in Chapter 4, several conclusions can be drawn for the utility of the proposed method and synthetic dataset while future extensions of this work are also suggested.

5.1 Conclusions

The first conclusion which can be derived is that the synthetic referring expressions generated with the proposed method can be effectively used to improve the performance of a deep neural network on the task of referring video object segmentation. The obtained results on different benchmarks for referring video object segmentation show that pre-training a model using the generated synthetic referring expressions, when it is additionally trained with human-produced referring expressions, increases its ability to generalize across different datasets.

Moreover, the experimental results show that the observed gains using the synthetic referring expressions are higher when the model is not fine-tuned on the human-produced referring expressions of the target dataset. What can be deduced from this finding is that, a large-scale dataset of synthetic referring expressions can be more useful in scenarios where training data for the target dataset are not available, which can be true for many real world applications where new data from different sources are seen at test time. This ability of applying a model trained on one source domain (*e.g.* one dataset) to another target domain (*e.g.* another dataset) is called domain adaptation and it is a field that has attracted much attention in the last years.

On the other hand, when directly comparing training a model with human-produced referring expressions versus training purely on synthetic referring expressions on the same videos, it is observed that human annotations yield better results. However, it is important to note that the proposed method requires no additional annotation effort whereas human annotations can be unattainable in many cases due to time or money constraints.

Finally, an ablation study concerning the information included in the synthetic referring expressions, confirms that attributes of objects, like their color, are important and can significantly improve the segmentation accuracy in the task of referring video object segmentation. This conclusion about the role of attributes in referring expressions is reported in previous works [Bell20] and the present work confirms the importance of attributes also in synthetic referring expressions.

5.2 Future Directions

As already mentioned in the previous chapters, the formulation of the proposed method for generating synthetic referring expressions allows its application to any other existing object detection or segmentation dataset since only object classes and bounding boxes are required. Thus, a possible

future direction would be to apply the proposed method to other datasets for pre-training the model, where RefCOCO could be an option. Since the best results, presented in Chapter 4, were obtained by pre-training the network on RefCOCO [Kaze14] (*i.e.* human-produced referring expressions) and then on the proposed synthetic referring expression for the videos of YouTube-VIS [Yang19], one possible future work could be to produce synthetic referring expressions for the images of the RefCOCO dataset. This way an annotation cost-free pre-training could be made as well as a study of the trade-off between the annotation cost and segmentation performance by using a variable ratio of human-produced to synthetic referring expressions.

Another possible direction would be to enhance the proposed method by adding more cues to the existing ones. An idea would be to use scene-graph generation models [Xu17, Tang20] in order to predict relationships between the annotated objects. Scene graph generation aims at understanding a visual scene through the detection of objects and the relationships between them by generating a visually-grounded scene graph where nodes represent objects and edges relationships between them. Thus, in the same way that the proposed method predicts a set of attributes for the target objects, such a model could also detect relationships between objects, which could allow the creation of better synthetic referring expressions by including the predicted relationships in them. An alternative which could also enrich the generated synthetic referring expressions could be to train the attribute detector network on a different dataset with a bigger set of annotated attributes. For example, the GQA dataset [Huds19] has a much bigger set of 501 attributes compared to the 201 of Visual Genome [Kris17] which was used for training the attribute detector network of the proposed method.

Bibliography

[Ande18] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould and Anton van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

[Anto15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick and Devi Parikh, "Vqa: Visual question answering", in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[Bell16] Sean Bell, C Lawrence Zitnick, Kavita Bala and Ross Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2874–2883, 2016.

[Bell20] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres and Xavier Giro-i Nieto, "RefVOS: A Closer Look at Referring Expressions for Video Object Segmentation", *arXiv preprint arXiv:2010.00263*, 2020.

[Cael17] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers and Luc Van Gool, "One-shot video object segmentation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 221–230, 2017.

[Chen17a] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[Chen17b] Liang-Chieh Chen, George Papandreou, Florian Schroff and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation", *arXiv preprint arXiv:1706.05587*, 2017.

[Chen19a] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen and Tyng-Luh Liu, "See-through-text grouping for referring image segmentation", in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7454–7463, 2019.

[Chen19b] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin and Ming-Hsuan Yang, "Referring expression object segmentation with caption-aware consistency", *arXiv preprint arXiv:1910.04748*, 2019.

[Cort95] Corinna Cortes and Vladimir Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[Dai16] Jifeng Dai, Yi Li, Kaiming He and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks", in *Advances in neural information processing systems*, pp. 379–387, 2016.

[Dala05] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection", in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[Das17] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh and Dhruv Batra, "Visual dialog", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335, 2017.

[Deng09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[Devl19] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[Doso15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers and Thomas Brox, "Flownet: Learning optical flow with convolutional networks", in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

[Elli13] Desmond Elliott and Frank Keller, "Image description using visual dependency representations", in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302, 2013.

[Erha14] Dumitru Erhan, Christian Szegedy, Alexander Toshev and Dragomir Anguelov, "Scalable object detection using deep neural networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147–2154, 2014.

[Frie18] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein and Trevor Darrell, "Speaker-follower models for vision-and-language navigation", in *Advances in Neural Information Processing Systems*, pp. 3314–3325, 2018.

[Fuku80] Kunihiko Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[Gavr18] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li and Cees GM Snoek, "Actor and action video segmentation from a sentence", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5958–5966, 2018.

[Girs14] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[Girs15] Ross Girshick, "Fast r-cnn", in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[Gius13] Alessandro Giusti, Dan C Cireşan, Jonathan Masci, Luca M Gambardella and Jürgen Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks", in *2013 IEEE International Conference on Image Processing*, pp. 4034–4038, IEEE, 2013.

[Goel18] Vikash Goel, Jameson Weng and Pascal Poupart, "Unsupervised video object segmentation for deep reinforcement learning", in *Advances in Neural Information Processing Systems*, pp. 5683–5694, 2018.

[Good14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, "Generative adversarial nets", in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[He16] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[He17] Kaiming He, Georgia Gkioxari, Piotr Dollár and Ross Girshick, "Mask r-cnn", in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[Hoch97] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[Hols90] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet and Ph Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform", in *Wavelets*, pp. 286–297, Springer, 1990.

[Hu16a] Ronghang Hu, Marcus Rohrbach and Trevor Darrell, "Segmentation from natural language expressions", in *European Conference on Computer Vision*, pp. 108–124, Springer, 2016.

[Hu16b] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko and Trevor Darrell, "Natural language object retrieval", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555–4564, 2016.

[Hu20] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang and Huchuan Lu, "Bi-Directional Relationship Inferring Network for Referring Image Segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4424–4433, 2020.

[Huan16] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra et al., "Visual storytelling", in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, 2016.

[Huan20] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu and Bo Li, "Referring Image Segmentation via Cross-Modal Progressive Comprehension", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10488–10497, 2020.

[Huds19] Drew A Hudson and Christopher D Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

[Jhua13] H. Jhuang, J. Gall, S. Zuffi, C. Schmid and M. J. Black, "Towards understanding action recognition", in *International Conf. on Computer Vision (ICCV)*, pp. 3192–3199, December 2013.

[Joac98] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features", in *European conference on machine learning*, pp. 137–142, Springer, 1998.

[John17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick and Ross Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

[Kaze14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten and Tamara Berg, "ReferItGame: Referring to Objects in Photographs of Natural Scenes", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, 2014.

[Khor18] Anna Khoreva, Anna Rohrbach and Bernt Schiele, "Video object segmentation with language referring expressions", in *Asian Conference on Computer Vision*, pp. 123–141, Springer, 2018.

[Khor19] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox and Bernt Schiele, "Lucid data dreaming for video object segmentation", *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1175–1197, 2019.

[Kris17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations", *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[Kriz12] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[LeCu90] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard and Lawrence D Jackel, "Handwritten digit recognition with a back-propagation network", in *Advances in neural information processing systems*, pp. 396–404, 1990.

[Li18] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen and Jiaya Jia, "Referring image segmentation via recurrent refinement networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2018.

[Lin14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C Lawrence Zitnick, "Microsoft coco: Common objects in context", in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[Lin17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan and Serge Belongie, "Feature pyramid networks for object detection", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[Liu16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C Berg, "Ssd: Single shot multibox detector", in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[Liu17] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu and Alan Yuille, "Recurrent multimodal interaction for referring image segmentation", in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1271–1280, 2017.

[Liu19] Runtao Liu, Chenxi Liu, Yutong Bai and Alan L Yuille, "Clevr-ref+: Diagnosing visual reasoning with referring expressions", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4185–4194, 2019.

[Long15] Jonathan Long, Evan Shelhamer and Trevor Darrell, "Fully convolutional networks for semantic segmentation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[Lowe99] David G Lowe, "Object recognition from local scale-invariant features", in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[Mans15] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba and Ruslan Salakhutdinov, "Generating images from captions with attention", *arXiv preprint arXiv:1511.02793*, 2015.

[Mao16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille and Kevin Murphy, "Generation and comprehension of unambiguous object descriptions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

[Marg18] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero and Pablo Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 630–645, 2018.

[Miko13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean, "Distributed representations of words and phrases and their compositionality", in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[Moga19] Aditya Mogadala, Marimuthu Kalimuthu and Dietrich Klakow, "Trends in integration of vision and language research: A survey of tasks, datasets, and methods", *arXiv preprint arXiv:1907.09358*, 2019.

[Noh15] Hyeonwoo Noh, Seunghoon Hong and Bohyung Han, "Learning deconvolution network for semantic segmentation", in *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

[Papa15] George Papandreou, Iasonas Kokkinos and Pierre-André Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 390–399, 2015.

[Peng15] Xingchao Peng, Baochen Sun, Karim Ali and Kate Saenko, "Learning deep object detectors from 3d models", in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1278–1286, 2015.

[Penn14] Jeffrey Pennington, Richard Socher and Christopher D Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[Pera16] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation", in *Computer Vision and Pattern Recognition*, 2016.

[Pete18] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer, "Deep Contextualized Word Representations", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018, Association for Computational Linguistics.

[Pont17] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung and Luc Van Gool, "The 2017 DAVIS Challenge on Video Object Segmentation", *arXiv:1704.00675*, 2017.

[Radf18] Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever, "Improving language understanding with unsupervised learning", *Technical report, OpenAI*, 2018.

[Redm16] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi, "You only look once: Unified, real-time object detection", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[Reit92] Ehud Reiter and Robert Dale, "A fast algorithm for the generation of referring expressions", in *COLING 1992 Volume 1: The 15th International Conference on Computational Linguistics*, 1992.

[Ren15] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", in *Advances in neural information processing systems*, pp. 91–99, 2015.

[Ronn15] Olaf Ronneberger, Philipp Fischer and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[Rume86] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams, "Learning representations by back-propagating errors", *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[Sale18] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson and Jose M Alvarez, "Effective use of synthetic data for urban scene semantic segmentation", in *European Conference on Computer Vision*, pp. 86–103, Springer, 2018.

[Seo20] Seonguk Seo, Joon-Young Lee and Bohyung Han, "URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. –, 2020.

[Serm13] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus and Yann LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks", *arXiv preprint arXiv:1312.6229*, 2013.

[Shi18] Hengcan Shi, Hongliang Li, Fanman Meng and Qingbo Wu, "Key-word-aware network for referring expression image segmentation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–54, 2018.

[Shri17] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang and Russell Webb, "Learning from simulated and unsupervised images through adversarial training", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107–2116, 2017.

[Silb18] Carina Silberer and Manfred Pinkal, "Grounding semantic roles in images", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2616–2626, 2018.

[Simo14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[Spec16] Lucia Specia, Stella Frank, Khalil Sima'an and Desmond Elliott, "A shared task on multimodal machine translation and crosslingual image description", in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 543–553, 2016.

[Tang20] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi and Hanwang Zhang, "Unbiased scene graph generation from biased training", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3716–3725, 2020.

[Tayl53] Wilson L Taylor, ""Cloze procedure": A new tool for measuring readability", *Journalism quarterly*, vol. 30, no. 4, pp. 415–433, 1953.

[Uijl13] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers and Arnold WM Smeulders, "Selective search for object recognition", *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[Varo17] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev and Cordelia Schmid, "Learning from synthetic humans", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 109–117, 2017.

[Vasw17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin, "Attention is all you need", in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[Vent19] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i Nieto, "Rvos: End-to-end recurrent network for video object segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5277–5286, 2019.

[Viol01] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features", in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, IEEE, 2001.

[Wang19a] Hao Wang, Cheng Deng, Junchi Yan and Dacheng Tao, "Asymmetric Cross-Guided Attention Network for Actor and Action Video Segmentation From Natural Language Query", in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3939–3948, 2019.

[Wang19b] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi and Haibin Ling, "Learning unsupervised video object segmentation through visual attention", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3064–3074, 2019.

[Wu20] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui and Subhransu Maji, "PhraseCut: Language-Based Image Segmentation in the Wild", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Xie19] Ning Xie, Farley Lai, Derek Doran and Asim Kadav, "Visual entailment: A novel task for fine-grained image understanding", *arXiv preprint arXiv:1901.06706*, 2019.

[Xing15] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", in *Advances in neural information processing systems*, pp. 802–810, 2015.

[Xu15] C. Xu, S.-H. Hsieh, C. Xiong and J. J. Corso, "Can Humans Fly? Action Understanding with Multiple Classes of Actors", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Xu17] Danfei Xu, Yuke Zhu, Christopher B Choy and Li Fei-Fei, "Scene graph generation by iterative message passing", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5419, 2017.

[Xu18] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen and Thomas Huang, "Youtube-vos: Sequence-to-sequence video object segmentation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 585–601, 2018.

[Yang19] Linjie Yang, Yuchen Fan and Ning Xu, "Video instance segmentation", in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5188–5197, 2019.

[Ye19] Linwei Ye, Mrigank Rochan, Zhi Liu and Yang Wang, "Cross-modal self-attention network for referring image segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10502–10511, 2019.

[Yu16] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg and Tamara L Berg, "Modeling context in referring expressions", in *European Conference on Computer Vision*, pp. 69–85, Springer, 2016.

[Yu18] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal and Tamara L Berg, "Mattnet: Modular attention network for referring expression comprehension", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018.

[Zeil11] Matthew D Zeiler, Graham W Taylor and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning", in *2011 International Conference on Computer Vision*, pp. 2018–2025, IEEE, 2011.