Panoptic Image Segmentation

Adepu Adarsh Sai AI20BTECH11001
Dhatri Nanda AI20BTECH11002
Savarana Datta Reddy AI20BTECH11008
Nelakuditi Rahul Naga AI20BTECH11029

Abstract

Panoptic segmentation unifies the typically distinct tasks of semantic segmentation (assign a class label to each pixel) and instance segmentation (detect and segment each object instance). Panoptic segmentation requires generating a coherent scene segmentation that is rich and complete, which is an important step towards real world vision systems.

1. Introduction

In the early days of computer vision, things i.e; objects such as people, animals, tools received the dominant share of attention. But studying systems that recognize stuff i.e; amorphous regions of similar texture or material such as grass, sky, road is also important. Studying things is typically formulated as the task of instance segmentation, where the goal is to detect each object and delineate it with a segmentation mask. In contrast, studying stuff is most commonly formulated as a task known as semantic segmentation. Panoptic segmentation refers to a unified, global view of segmentation i.e; it encompasses both thing and stuff classes. In Panoptic segmentation, each pixel of an image is assigned a semantic label and an instance id. Pixels with the same label and id belong to the same object; for stuff labels the instance id is ignored.

2. Problem Statement

Designing an image encryption and decryption system using panoptic segmentation for securing confidential information. The system involves segmenting the image using panoptic segmentation to identify the target region, encrypting the desired message within the target region, and then creating a new image with the encrypted message. The system should be able to decrypt the message using the same panoptic segmentation process and retrieve the original message. The objective of this project is to develop an efficient and secure method for protecting sensitive information within an image using advanced image segmentation techniques.

3. Image Segmentation

Image segmentation is the process of splitting a digital picture into several image segments, also known as image regions or image objects, in digital image processing and computer vision (sets of pixels). The purpose of image segmentation is to simplify and transform an image's representation into something more relevant and easier to examine. Image segmentation is commonly used to find objects and boundaries (lines, curves, and so on) in pictures. Image segmentation is the process of labelling every pixel in a picture so that pixels with the same label have specific properties.

3.1. Types of Image Segmentation

1. **Semantic Segmentation:** It is a method for determining the object's class for each pixel. For example, in a figure, all cars on the road are segmented as one item, while the background is segmented as one object.

2. **Instance Segmentation:** It is a method for identifying a corresponding instance of an item for each pixel. It recognises each individual object of interest in the picture. For example, in a figure, each car on the road is segmented as a unique entity.

3. **Panoptic segmentation:** It combines semantic segmentation with instance segmentation. Panoptic segmentation, like semantic segmentation, determines the class to which each pixel belongs. Panoptic segmentation, as distinct to semantic segmentation, distinguishes across various instances of the same class.

4. Literature Review

This paper [1] discusses about a weakly supervised model that jointly performs non-overlapping semantic and instance segmentation for both thing and stuff classes. They address the problems of semantic and instance segmentation using only weak annotations in the form of bounding boxes for thing classes and image-level tags for stuff classes, which is cost-effective. The model described in this paper achieves about 95% of fully-supervised accuracy on both semantic and instance segmentation tasks

on the Pascal VOC dataset. This emphasises that accurate models can be learned from only bounding box annotations, which are significantly quicker and cheaper to obtain than pixel-wise annotations. On Cityscapes dataset, using image-level annotations for stuff classes, they obtain 88.8% and 85.6% of fully-supervised performance for semantic segmentation and instance segmentation tasks respectively. The weakly-supervised model also provides better results than some of the state-of-the-art fully-supervised instance segmentation algorithms. The case where there are a mixture of weak and fully-labelled annotations, also known as Semi-supervised case is also described in the paper.

This paper [2] proposes a novel end-to-end Occlusion Aware Network (OANet) for panoptic segmentation, which can efficiently and effectively predict both the semantic and instance segmentation in a single model/network. It also introduces a novel spatial ranking module to deal with the occlusion problem between the predicted instances. A large receptive field can help the spatial ranking module get more context features and better results. The proposed model provides promising results on the COCO Panoptic benchmark dataset.

The task of instance-aware semantic segmentation is covered in this [3] research. The main objective is to provide a straightforward way using a novel modelling paradigm, which differs from existing approaches in how advantages and disadvantages are traded off. This method, known as InstanceCut, illustrates the issue using two different output modalities: (i) a semantic segmentation that is instance-agnostic and (ii) all instance borders. The former is obtained from a novel instance aware edge detection model, while the later from a conventional convolutional neural network for semantic segmentation. The authors combine these two modalities into a novel MultiCut formulation in order to reason broadly about the best way to divide an image into instances. The method is evaluated using the difficult CityScapes dataset. Despite the approach's conceptual simplicity, it produces the best results of all approaches that have been published, and it performs especially well for rare object classes.

A Bayesian framework that uses Bayesian probability theory for parsing images into their constituent visual patterns is proposed in this paper [4]. The framework gives a rigourous way to combine segmentation with object detection and recognition. The parsing algorithm outputs a scene representation as a parsing graph, similar to parsing sentences in speech and natural language. The proposed computational framework integrates generative and discriminative methods of inference, both of which are extensively used by the vision and machine learning communities. They also investigate visual patterns such as texture and shading, and object patterns such as human faces and text. These types of patterns compete and cooperate to explain the image and so image parsing unifies image segmentation, object detection, and recognition.

In this [5] they built their own dataset for training the models. The **Mapillary Vistas Dataset** is a unique, large-scale street-level picture collection that contains 25,000 high-resolution photographs that have been classified into 66 item categories with extra, instance-specific labels for 37 classes. Annotation is done in a dense and fine-grained manner, with polygons used to define distinct objects. Their collection is 5 times larger than the entire number of fine annotations for Cityscapes and comprises photographs from all around the world, collected at diverse weather, season, and daytime circumstances. Pictures are captured using various imaging devices (mobile phones, tablets, action cameras, professional capture rigs) and by photographers with varying levels of competence. Their dataset has been constructed and produced in such a way that it covers diversity, richness of detail, and geographic scope. They propose semantic picture segmentation and instance-specific image segmentation as default benchmark tasks, with the goal of greatly furthering the development of state-of-the-art algorithms for visual road-scene interpretation.

Instance-specific semantic segmentation adds the complexity of recognising the pixels that form each object instance, thereby merging semantic segmentation with fine-grained object recognition.

The next interesting question would be, "What is the next frontier in visual recognition?". In [6], they provided one possible response to this question. They developed a thorough picture annotation that collects information beyond visible pixels and necessitates significant thinking about the overall structure of the scene. They specifically construct an amodal segmentation of each image: the whole extent of each region, not only the visible pixels, is noted. Annotators construct a partial depth order and highlight and label all conspicuous locations in the picture. As a result, the scene structure is rich, including viewable and occluded regions, figure-ground edge information, semantic labels, and object overlap.

For semantic amodal segmentation in [6], they generated two datasets. Initially, they classify 500 photos from the BSDS dataset with numerous annotators per image, allowing them to analyse human annotation statistics. They demonstrated that the suggested full-scene annotation is remarkably consistent among annotators, even for regions

and edges. Second, they annotated 5000 COCO pictures. Using a bigger dataset, they were able to test a variety of algorithmic methods for amodal segmentation and depth sorting. They defined tangible new challenges for the community by introducing unique measures for these activities and defining solid baselines.

For the problem, Object instance segmentation, we have a method proposed in [7] called Sequential Grouping Networks SGN. Object instance segmentation is a highly complex task comprising of object detection, semantic segmentation and occlusion. Works related to instance segmentation focused on classifying and labelling the object. The task of object segmentation is divided into subtasks, predicting object breakpoints(horizontal and vertical), then creating horizontal and vertical line segments, then grouping these into connected components and then merging these objects into coherent object instances. Each of the above sub-tasks has been explained in the paper using different methods. To test the learned method in the paper, the method is evaluated on two datasets. For the city scapes dataset containing 5,000 images, the method was seen to have outsmarted other existing methods (this dataset is generally seen as challenging).

An approach to holistic scene understanding is discussed in this [8] paper. This paper discusses combining the sub-tasks like scene classification, image labelling, object detection, semantic regions present in the scene and other multiple related aspects to understand the scene more deeply. The holistic scene approach respects boundaries well so we can represent a problem with few variables(representing class labels). When tested on datasets, this joint model improved the accuracy of scene classification as well as segmentation and object detection. For this holistic task, a structure prediction framework is used to learn the weights by defining a holistic loss function. There are Object Reasoning Potentials and Class Presence Potentials defined as part of this method. This is then evaluated for both semantic segmentation and object segmentation on two different datasets. And they have shown that this model boosts both scene classification and object detection along with improving segmentation accuracy along with following the boundaries. This approach achieves state-of-the-art performance.

5. Panoptic Segmentation

This paper [9] proposes a panoptic quality (PQ) metric that can measure the performance of both "stuff" and "things" categories in a way that is unified. As seen earlier, panoptic segmentation(PS) is the unification of semantic segmentation and instance segmentation(object detection). And this paper aims to introduce a uniform evaluation metric whose task is to encompass both stuff and thing classes. Each pixel is assigned an instance id and a semantic label. The format for panoptic segmentation:

1. $\mathcal{L} := \{0, ..., L-1\}$ is the predetermined set of L semantic classes.
2. the label subsets corresponding to the stuff and thing are as $\mathcal{L}^{St}$ and $\mathcal{L}^{Th}$,

Instance segmentation allows overlapping segments but panoptic segmentation allows us to assign only one instance id and one semantic label to each pixel. So, for PS, no overlaps are possible. The following are the essential characteristics that a suitable measure for PS should possess: "Completeness, Interpretability and Simplicity".

The first step is segment matching. When given a predicted and ground truth panoptic segmentation of an image, a theorem is proposed saying that the former and latter can be matched only if their IoU (intersection over union) is strictly greater than 0.5. This can be seen based on some simple calculations as follows:

$$IoU(p_i, g) = \frac{|p_i \cap g|}{|p_i \cup g|} \leq \frac{|p_i \cap g|}{|g|} \quad \forall\, i \in \{1, 2\}$$

where g is the ground truth segment and p_1 and p_2 are two predicted segments. We use the definition $p_1 \cap p_2 = \phi$ and $|p_i \cap g| \geq |g|$ for the above result.
Then scanning over all i, we get

$$IoU(p_1, g) + IoU(p_2, g) \leq \frac{|p_1 \cap g| + |p_2 \cap g|}{|g|} \leq 1$$

because $|p1 \cap g| + |p_2 \cap g| \leq g$.

The advantage of the above theorem is that it is easy and effective because the matches are unique.
Note: For smaller IoU, other matching techniques will be required.

The next step is PQ computation. It is given below

$$PQ = \frac{\sum_{(p,g)\in TP} IoU(p, g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

where TP, FP, and FN are the true positives, false positives and false negatives sets to which the unique matching splits predicted and ground truth segments of each class.

Further,

$$PQ = \underbrace{\frac{\sum_{(p,g)\in TP} IoU(p, g)}{|TP|}}_{\text{segmentation quality(SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality(RQ)}}$$

When written this way, though the two variables are not independent, it's easier to analyse and estimate. Void labels are removed in the process. Even if the outputs contain void pixels, the evaluation is not affected.

Finally, panoptic segmentation treats all classes in a unifrom way and unifies evaluation over all classes.

6. DETR

This paper [10] proposes a method that views object detection as a direct set prediction problem and uses a framework called as DEtection TRansformer or DETR. It is a set-based global loss that forces unique predictions via bipartite matching, and a transformer encoder-decoder architecture. DETR can be easily generalized to produce panoptic segmentation in a unified manner. For direct set predictions in detection, the model uses two essential ingredients :

1. A set prediction loss that enforces unique matching between predicted and ground truth boxes.

2. An architecture that predicts a set of objects and models their relation in a single pass.

Let y be the ground truth set of objects, and $\hat{y} = \{\hat{y_i}\}_{i=1}^{N}$, be the set of N predictions. To find a bipartite matching between these two sets we search for a permutation of N elements $\sigma \in S_N$ with the lowest cost :

$$\hat{\sigma} = \underset{\sigma \in S_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

where $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is the pairwise matching cost between ground truth y_i and a prediction with index $\sigma(i)$. Each element i of the ground truth set can be seen as $y_i = (c_i, b_i)$, where c_i is the target class label and $b_i \in [0, 1]$ is a vector that defines ground truth box center coordinates and its height and width relative to the image size. For the prediction with index $\sigma(i)$ we define probability of class c_i as $\hat{p}_{\sigma(i)}(c_i)$ and the predicted box as $\hat{b}_{\sigma(i)}$. We now have :

$$\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = -\mathbf{1}_{\{c_i \neq \phi\}} \hat{p}_{\sigma(i)} + \mathbf{1}_{\{c_i \neq \phi\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)})$$

where $\mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)})$ is the loss that scores the bounding boxes. The second step is to compute the loss function, the Hungarian loss ($\mathcal{L}_{\text{H}}$) for all the pairs matched in the previous step as follows :

$$\mathcal{L}_{\text{H}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbf{1}_{\{c_i \neq \phi\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}(i)}) \right]$$

The overall DETR architecture contains three main components - A CNN backbone to extract a compact feature representation, an encoder-decoder transformer, and a simple feed forward network (FFN) that makes the final detection prediction.

DETR can be naturally extended to solve the task of panoptic segmentation by adding a mask head on top of the decoder outputs. First, DETR is trained to predict boxes around both stuff and things classes on COCO dataset. We also add a mask head which predicts a binary mask for each of the predicted boxes. It takes as input the output of transformer decoder for each object and computes multi-head attention scores of this embedding over the output of the encoder. To predict the final panoptic segmentation, we simply use an argmax over the mask scores at each pixel, and assign the corresponding categories to the resulting masks.

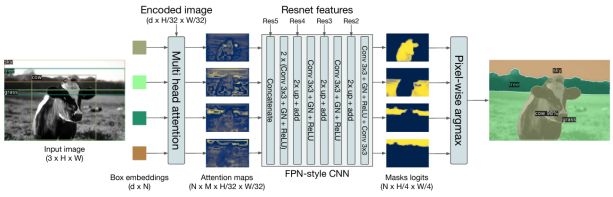

Figure 1. DETR Architecture

7. Fast R-CNN's

This paper [11] introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a kind of fully convolutional network (FCN) that simultaneously predicts object bounds and objectness scores at each position. It can be trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. Faster R-CNN is composed of two modules, the first module is a deep fully convolutional network that proposes regions, and the second module is the Fast R-CNN detector that uses the proposed regions. The entire system is a single, unified network for object detection.

A Region Proposal Network (RPN) takes an image as input and outputs a set of rectangular object proposals, each with an objectness score. For training RPNs, we assign a binary class label to each anchor. The loss function for an image is defined as follows :

$$\mathcal{L}(\{p_i\}, \{t_i\}) = \frac{1}{N_{\text{cls}}} \sum_{i} \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \frac{\lambda}{N_{\text{reg}}} \sum_{i} p_i^* \mathcal{L}_{\text{reg}}(t_i, t_i^*)$$

where i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The ground-truth label is p_i^*. t_i and t_i^* are vectors representing

the four parameterized coordinates of the predicted bounding box and ground-truth box respectively. $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{reg}}$ denote the classification loss and the regression loss respectively. λ is a balancing parameter. Finally, for the detection network, various algorithms that learn a unified network composed of RPN and Fast R-CNN with shared convolutional layers are proposed.

8. An introduction to steganography methods

Stenography [12] vaguely means embedding secret data in a cover i.e., protecting a message. In this process, there are two types of materials used: message and carrier. Let's look into different stenography techniques

1. **Text Stenography**: This involves text formatting or characteristics to hide information while keeping it difficult to notice by the reader. The challenge in designing such techniques lies in achieving a balance between reliable decoding and minimal detectable change. There are three coding techniques introduced:

   Line-Shift Coding is a technique that involves vertically shifting lines in a document. Here, decoding can be done without the original image in some cases.

   Word-Shift Coding is another technique that involves horizontally shifting words within text lines in a document. It is only applicable to documents with variable spacing between adjacent words.

   Feature Coding is a technique that examines chosen text features and alters them based on a codeword. It requires an original image for decoding and typically focuses on vertical endlines altering their length without changing the endline feature.

2. **Image Steganography** means hiding information inside images. There are very limited practical applications to this. To hide a message in an image, modifications are done in noisy areas with colour variations by using the below techniques:

   Least Significant Bits is a method which involves modulating the least significant bit of the image's pixels. The use of lossless compression format is necessary to prevent loss of hidden information.

   Masking and filtering techniques involve modifying the visible properties of an image by creating markings or changing the luminance of certain parts of the image. The information here is hidden inside the visible part of the image and not at the noise level.

   Transformations involve using and modifying DCT to hide a secret inside an image. This method is more complex than LSB and masking techniques.

3. **Audio Steganograpy** embeds a secret message into a digitized audio signal.

   Spread Spectrum is a modulation technique used in telecommunications, with two approaches Direct Sequence Spread Spectrum(DSSS) and Frequency Hopping Spread Spectrum(FHSS). DSSS multiplies the data being transmitted by a pseudorandom noise signal, while FHSS pseudo-randomly retunes the carrier.

   Echo Hiding involves embedding a secret message as an echo in a cover audio signal. This technique is also applicable to video files, where a video sequence is used as cover media, and a secret key can be used during the embedding process to produce a "stego-video". At the receiving end, the secret key is used to extract the secret message from the stego-object.

   There are a few other techniques used in audio steganography like **LSB Coding** and **Phase Coding**

9. Coverless Image Steganography

In [13], a novel coverless image steganography method based on image segmentation is proposed. They use ResNet to extract semantic features, and segment the object areas from the image (from COCO and VOC datasets) through Mask RCNN for information hiding. These selected object areas have ethical structural integrity which help reduce the information loss due to malicious attacks. In Coverless Steganography, the hiding process is implemented by finding an image or text that already contains the secret information.

In the proposed steganography scheme, first of all a large number of object areas are segmented through Mask RCNN and the needed object areas are selected. The areas are selected by setting a threshold, thus giving areas which can escape geometric attacks. Then a sequence of the object areas is generated using robust hash algorithms and an index is established for feature matching. Stego-images are then matched and transmitted to the receiver. The receiver uses Mask RCNN to obtain the object areas from stego-images according to feature points. Then, the sequences of the object areas are generated by hash algorithm and sequentially concatenated to obtain secret information. The process is depicted below :

The robustness of the algorithm i.e; the ability of the the encoded information to resist these attacks is determined by the success rate of secret data extraction. It is calculated as follows :

$$SR = \frac{\sum_{i=1}^{m} h_i \cap e_i}{m}$$

where where m represents the number of transmitted object areas, h_i represents the hided bits of each area and e_i rep-

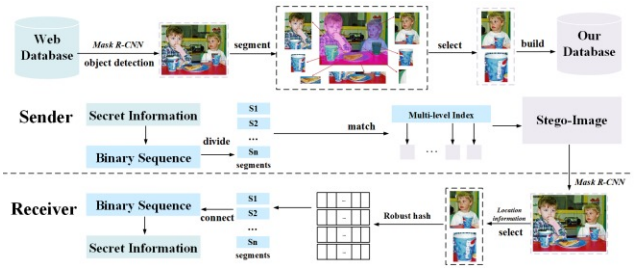

Figure 2. Coverless Steganography using Image Segmentation

resents the corresponding extracted bits. The method exhibited has a significant advantage when being geometric attacked, which ensures the security of secret information.

10. Final Results

This study employs the DETR algorithm for panoptic segmentation of images. We applied the DETR (DEtection TRansformer) model for panoptic image segmentation. The model was trained on the COCO (Common Objects in Context) dataset, which contains over 330k images with 80 object categories. (References - [14], [15])

(a) Original Image

(b) Panoptic Segmented Image

The project focuses on selecting a specific segment of an image and encrypting a message within it. The encrypted segment replaces the original segment in the image.

Figure 4. Targetted segment for encryption

We also wrote a decoding function to retrieve the encrypted message from the selected segment. We used *lib* function from the *stegano* library for encrypting and decrypting the data.

Figure 5. Image after encryption

Additionally, we demonstrated the encryption and decryption of an entire image using a similar approach. The scope of this project is limited to the use of DETR for image segmentation and message encryption within specific image segments.

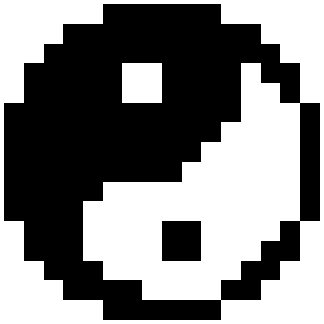

Figure 6. Image that is encrypted

11. Future Work

1. **Improve Encryption Algorithm**: Investigating and creating more complex encryption algorithms that can provide higher security assurances is one possible avenue for future research. This might involve investigating cutting-edge encryption methods like homomorphic encryption, which enables computations to be performed on encrypted data without first decrypting it. In addition, coverless picture steganography may be investigated as a better encryption method. With the use of a technique called coverless picture steganography, a message may be concealed within an image without the need for a separate cover image, mak-

ing it more resilient against assaults that depend on the recognition of a cover image. The system's security and dependability might be increased by researching and utilising coverless picture steganography techniques.

2. **Integration with existing systems**: This system might need to be integrated with current systems or procedures in actual applications. Future development may concentrate on creating APIs or integration interfaces that enable the system to be utilised in combination with other devices or programmes.

3. **Video encrypting**: Another potential direction is to extend the system to work with the videos. This could involve encrypting a sequence of data in different frames of a particular object in a video.

References

[1] Qizhu Li, Anurag Arnab, and Philip H.S. Torr. Weakly and semi-supervised panoptic segmentation, 2018. University of Oxford. 1

[2] H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, and W. Jiang. An end-to-end network for panoptic segmentation, 2019. Zhejiang University, Megvii Inc. , Huazhong University of Science and Technology, Peking University, The University of Tokyo. 2

[3] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut, 2016. TU Dresden, Dresden, Germany and MPI for Informatics, Saarbrucken, Germany. 2

[4] Z. Tu, X. Chen, A. L. Yuille, and S.C. Zhu. Image parsing: Unifying segmentation, detection, and recognition, 2005. IJCV. 2

[5] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bul'o, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes, 2008. ICCV. 2

[6] Y. Zhu, Y. Tian, D. Mexatas, and P.Dollar. Semantic amodal segmentation, 2017. CVPR. 2

[7] S. Liu, J. Jia, S. Fidler, and R. Urtasun. Sgn: Sequential grouping networks for instance segmentation, 2017. CVPR. 3

[8] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation, 2012. CVPR. 3

[9] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation, 2019. Facebook AI Research (FAIR) and HCI/IWR, Heidelberg University, Germany. 3

[10] Nicolas Carion, Francisco Massa, Nicolas Usunier Gabriel Synnaeve, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. Facebook AI. 4

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 4

[12] Masoud Nosrati, Ronak Karimi, and Mehdi Hariri. An introduction to steganography methods, 2011. WAP Journal. 5

[13] Yuanjing Luo, Jiaohua Qin, , Xuyu Xiang, Yun Tan, Zhibin He, and Neal N. Xiong. Coverless image steganography based on image segmentation, 2020. CMC. 5

[14] DETR Facebook Research. 6

[15] DETR Panoptic Demo. 6