

Investigating transformers in the decomposition of polygonal shapes as point collections

Andrea Alfieri, Yancong Lin, and Jan C. van Gemert

Computer Vision Lab, Delft University of Technology, the Netherlands

Abstract

Transformers can generate predictions in two approaches: 1. auto-regressively by conditioning each sequence element on the previous ones, or 2. directly produce an output sequences in parallel. While research has mostly explored upon this difference on sequential tasks in NLP, we study the difference between auto-regressive and parallel prediction on visual set prediction tasks, and in particular on polygonal shapes in images because polygons are representative of numerous types of objects, such as buildings or obstacles for aerial vehicles. This is challenging for deep learning architectures as a polygon can consist of a varying cardinality of points. We provide evidence on the importance of natural orders for Transformers, and show the benefit of decomposing complex polygons into collections of points in an auto-regressive manner.

1. Introduction

Predicting polygonal shapes in images is a high-level task that is representative of many vision problems. One example is to automatically extract vectorized building outlines from high-resolutions satellite images [11, 12, 16, 20]. Similarly, polygon detection can be beneficial for vision-based flight control of unmanned aerial vehicles (UAVs), as demonstrated in the Autonomous Drone Racing [17], where drones fly autonomously through a sequence of polygon-shaped gates purely relying on vision signals. The performance of polygon detection is critical for UAVs as it directly affects the precision of navigation.

In this work, we treat polygonal shape prediction as a *collection prediction* task such as sequences and sets. Many kinds of data can be naturally represented using collection and many machine learning tasks can be viewed as a collection prediction problem, such as predicting the collection of points of a polygon [36], detecting objects in an image [4, 35], estimating the pose of humans by detecting a set of

key-points [3, 29], or predicting multiple labels for the same sample [26]. The difficulty in such problems arises for two main reasons: because the *cardinality* (i.e. the number of elements) of the collection is unknown and can vary among different samples, and because collections when treated as a set are *permutation-invariant*. However, canonical convolutional neural networks are not able to tackle these problems by design [19].

Transformers [30] have achieved impressive results on numerous collection prediction tasks [4, 19], thanks to the attention module [1], a solution capable of aggregating information from the entire collection while also satisfying permutation invariance.

The work in [30] first presents Transformers as an auto-regressive sequence-to-sequence model that generates a collection of output tokens one by one. Numerous following works [5, 10, 14] introduce new variants of Transformers to reduce compute latency. However, the vast majority of auto-regressive Transformers focuses on sequential tasks only, such as machine translation [10, 14] or speech recognition [5]. Recently, parallel Transformers have successfully been applied into computer vision, especially in object detection [4, 37], where the model outputs a set of bounding boxes *in parallel*. However, parallel Transformers rely on computationally intensive strategies, such as oversampling and Hungarian matching [18].

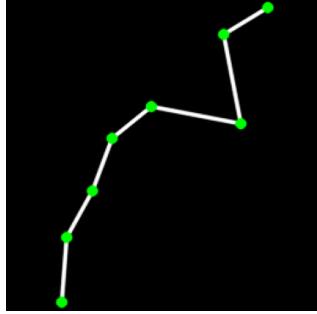
The difference between auto-regressive and parallel Transformers is fundamental. Let us consider an image containing a collection of three objects (namely A , B and C) which we are trying to detect. When asking a parallel Transformer to learn this task, we are actually asking the model to learn the joint probability of these three variables, conditioned on the model parameters Θ , namely:

$$P(A, B, C \mid \Theta) \quad (1)$$

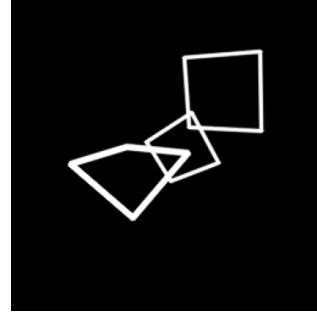
In contrast, when the auto-regressive approach is presented with the same task, what it needs to learn is the chain of



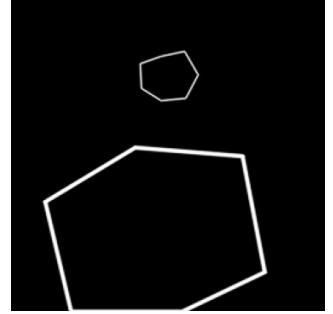
(a) Points



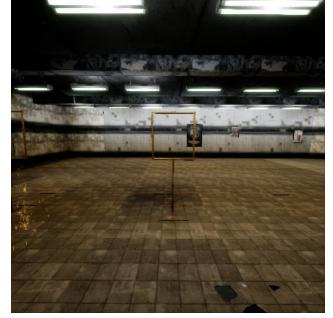
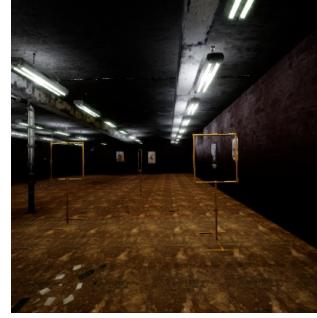
(b) Lines



(c) Gates



(d) Polygons



(e) Synthetic gate dataset.

Figure 1: Examples of point collections used in our paper.

conditional probability distributions defined by:

$$P(A | \Theta) \cdot P(B | A, \Theta) \cdot P(C | A, B, \Theta) \quad (2)$$

In the second case, the Transformer is asked to produce one element of the collection at a time, by exploiting information about the previous predictions. This is similar to machine translation, in which the Transformer outputs one word, conditioned on the previously generated output. Eq. (1) and Eq. (2) are equal by definition of the general product rule of probability, and Transformers are capable of modelling both. However, using Eq. (2) with Transformers imposes an order even when there might not be a natural order present.

In this work, we study the difference between auto-regressive and parallel Transformers on polygonal shape prediction viewed as a collection of points, including individual points, lines, gates and polygons, as shown in Fig. 1. Our contributions are: (1) We test auto-regressive and parallel models on four collection prediction datasets and one sequential dataset to provide the reader with a full picture of the advantages and disadvantages of both approaches. (2) We show that the conditional decomposition of the collection can be beneficial for Transformers when there is an explicit order in the elements of a set, such as predicting a collection of points on a line. (3) we show empirically that the conditional decomposition benefits from a particular or-

der than others on a polygon dataset.

2. Related work

Polygonal shape detection: The work in [12] views satellite image mapping as a polygon prediction task and makes use of a convolutional network to output the polygon vertices directly. The work in [20] first detects bounding boxes of buildings, and then uses recurrent networks to extract vectorized building footprints. The work in [11] proposes to detect frame field for constructing building topology via fully convolutional networks and is able to precisely segment aerial images. Different from these works, we study the effect of self-attentions or Transformers in polygonal shape prediction.

Object detection as set prediction. DETR [4] first presents object detection as a set prediction task, and removes the need for non-maximal suppression on a large amounts of *anchor boxes*. Similar to DETR, we also consider polygonal shapes in an image as a set. Different from DETR which only output bounding boxes, we produce vectorized polygons as a collection of vertices.

Transformers. Transformers were originally introduced by [30] as a novel auto-regressive, sequence-to-sequence model, and gained popularity thanks to their ability to dispense entirely with recurrence and support parallel processing of sequences. Their stunning results on machine trans-

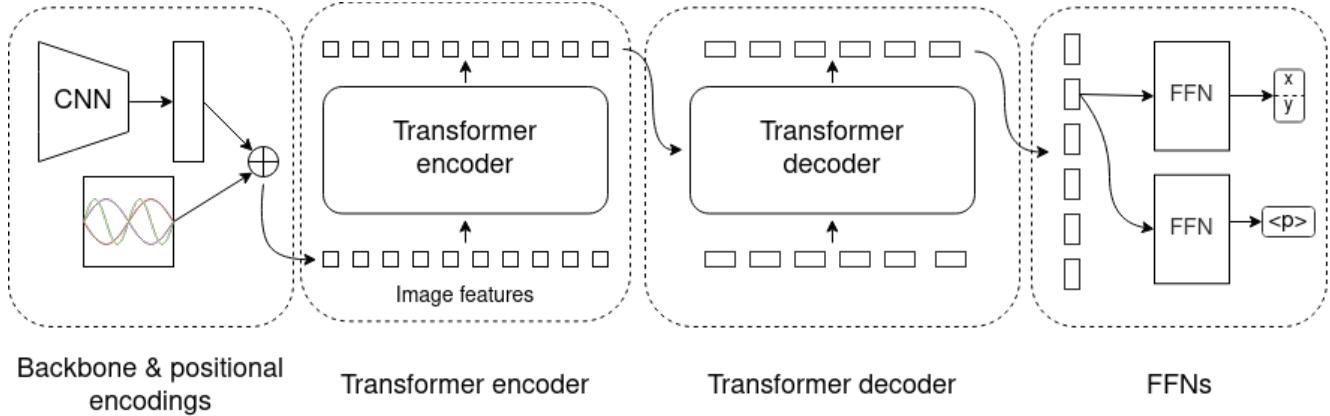


Figure 2: **The auto-regressive model for predicting polygons.** This model exploits the *sentences of tokens* idea [30] which generates tokens auto-regressively. There are only four possible tokens in its vocabulary: *start*, *point*, *end-of-polygon* and *end*. The feed-forward network (FFN) at the end of the pipeline produces a class label and a pair of coordinates (x, y).

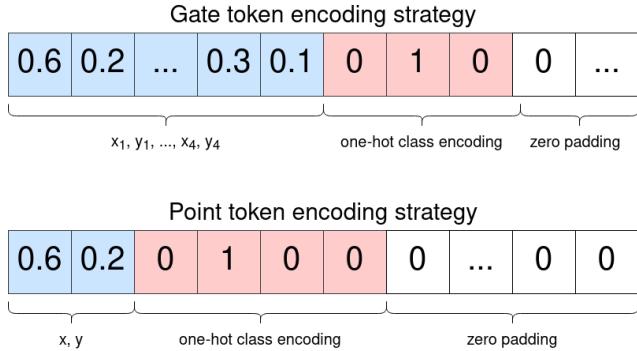


Figure 3: **Token embeddings for predicting gates and polygons.** The first $2n$ elements (in blue) of the vector represent the (x, y) coordinates of the n vertices/points, normalized by the image width and height. The following elements (in red) represent the class label in one-hot format. For gate detection, we have three red blocks because we have two special tokens *start* and *end*, and an object token *gate*. For polygon detection we have 4 red blocks because all possible tokens are *start*, *end*, *point* and *end-of-polygon*, as explained in 4.2. The vector is padded with zeros to reach the required dimensionality of 256.

lation and other language tasks [2, 7, 21, 24] have recently shed a light on employing Transformers for computer vision tasks, such as image recognition [8], object detection [4], segmentation [33], set prediction [19, 31] and other visual tasks [6, 13, 28, 32]. Inspired by these works, we study both auto-regressive and parallel Transformers on polygonal shape prediction.

Transformers for set prediction. Deep Sets [34] has proven that transforming all elements of an input set into some latent representations and then combining them through a permutation invariant function is a universal approximator of any set function. The work in [31] shows that the attention layer of Transformers can also be viewed as a generalization of the sum operation of Deep Sets and is therefore also a universal approximator of set functions. Moreover, Lee *et al.* [19] proposes an attention-based permutation-invariant framework which demonstrates superior results on set-structured data. Inspired by these works, we also study Transformers for set prediction task, but on a particular type of set data - polygonal shapes.

3. Datasets

To illustrate our setting, we first show examples of our datasets in Fig. 1. We use 4 toy datasets and 1 synthetic dataset. The toy images are manually generated during training and if not specified otherwise, we generate 3 million images for training and 10,000 images for testing.

The **point** dataset contains images with n white points randomly distributed over the image space. Each point's size is uniformly sampled from three possible values. The task is to predict the x, y coordinates of all points in any order. This dataset is an instance of a pure set prediction problem, with points representing the set elements.

The **line** dataset contains images picturing a single white line composed by 7 segments. A green point of fixed size is placed on top of each end of a segment. The line is generated by first drawing a straight line going from the bottom left corner of the image to the top right corner, and then randomly shifting 8 equally distributed points perpendicularly to the line direction by $r \in [-15\%, 15\%]$ with respect to

the image size. The task is to predict the x, y coordinates of the 8 points following the line order, starting from the end on the bottom left of the image. The set elements are represented in this case by the green points, and this dataset is an example of a set prediction task where there exists an explicit order in which we need to predict its elements.

The **gate** dataset contains images with n convex polygons of 4 corners, called *gates*. Each gate is generated by defining 4 equally spaced points on a circumference of random radius $r \in [5\%, 40\%]$ with respect to the image size. Each point is then shifted randomly in the direction of the radius and the four points are finally connected to define the gate. All four edges of the gates are of the same thickness, uniformly sampled out of 3 possible choices.

The **synthetic gate** dataset [9] contains realistic synthetic images generated with a graphical engine, which simulate the flight of a drone in different environments containing empty wireframe objects (EWFOs or *gates*). This dataset is a simulation of the images that a UAV would face in the IROS 2018 Autonomous Drone Race [17]. The training set contains 26,000 images from two different scenes, with light coming from outside and inside the rooms. The test set contains 3,000 images from two additional scenes with light coming only from artificial sources placed inside the rooms. For all scenes, different walls and pavement textures are used, as well as different artificial shapes and light intensities for internal lamps. All images contain 1 to 4 gates. For both synthetic and toy gate datasets, the task is to predict the position of the four corners in the image for each gate. These two datasets represent a set prediction scenario closer to real-life problems, where the complexity of the single set element prediction (the gate) is greater than a simple point prediction. The synthetic gates dataset increases the complexity even further by picturing gates with different backgrounds and different lighting situations.

Finally, the **polygon** dataset contains images with n polygons of $m \in [3, 7]$ corners, generated using the same technique as the gate dataset. As multiple non-convex polygons can be represented by the same set of points, the task is to predict the m corners of a polygon in clockwise order. Any starting point is accepted, and distinct polygons can be predicted in any order. This dataset represents the hardest set prediction task in which we require our models to predict a *set of ordered sets*.

For all datasets, n is a parameter varied for different experimental settings. All samples in the toy datasets are RGB images of size 256x256, while images from the synthetic gate dataset are of size 400x400. The coordinates of the labels are represented as the relative height/width to the image size with values in $[0, 1]$.

4. Models

All of the models that we implemented and tested are derived from DETR [4], an end-to-end Transformer which achieves competitive results against Faster R-CNN [25] on object detection. It takes advantage of a CNN backbone and parallel encoding-decoding Transformers to solve object detection as a set prediction task. In short, DETR is composed of four modules: a CNN backbone, a Transformer Encoder, a Transformer Decoder, and a feed-forward network (FFN). An example of our models is shown in Fig. 2. It pictures the auto-regressive variant used for the polygon toy setting dataset.

DETR takes as input an RGB image and extracts a high-level feature map of shape $C \times H \times W$, where C is the size of number of channels, and H and W are the spatial dimensions. The feature map is supplemented with fixed positional embeddings before the Transformer encoder. The Transformer encoder is a stack of 6 self-attention mechanisms, each of which consists of a standard self-attention layer followed by a feed forward network, a residual connection and layer normalization [4]. All of the models we tested are identical up to this stage of the architecture, but differ in the subsequent modules.

4.1. Parallel models

The design of parallel models is identical to DETR [4]. The N learned *object queries* are converted *in parallel* into N output embeddings of size 256 by the Transformer decoder, which are subsequently fed into a simple feed forward network (FFN) for classification and position regression.

Depending on the task, we modify the dimensionality of the FFN for position regression accordingly. For example, on the point and the line datasets, the output is a vector of size 2 representing the (x, y) coordinates of a point normalized by the image width and height, while the output on the gate dataset is an 8-element vector indicating the four vertices of a gate in clock-wise order. The class labels for all datasets are the same, namely the *object* class and the *no-object* class.

On the polygon dataset, the FFN is replaced by a simple multi-layer Elman RNN [27], since the output dimension is also a variable. The RNN model takes as input the embeddings from the *object* class and generates a sequence of points one by one. The RNN is possibly the minimal modification we could make to the architecture to work with polygons and have the smallest impact on the model’s properties and our experiments.

4.2. Auto-regressive models

The auto-regressive models diverge from the parallel ones on all set prediction tasks, because they work with *sentences of tokens* that are generated one by one, by conditioning the next prediction on the previous ones. In particular, each token is a vector of dimension 256 that represents an element of the set, or acts as a special representation. Special tokens can be:

- *start* or **S**: This token is at the beginning of all sentences and defines the starts of the computation.
- *end* or **E**: This token is at the end of all sentences and terminates the computation.
- *end-of-polygon* or **EOP**: This token separates polygons from polygons inside the same image. It acts similarly to the *period* token used in machine translation to separate words belonging to different sentences.

Depending on the task, object tokens can be:

- *point* or **P**: On the point dataset and the line dataset, each point is a 2-element vector representing the position of the (x, y) coordinate normalized by the image width and height.
- *gate* or **G**: On both toy and synthetic gate datasets, a gate is represented as a vector of size 8, which defines the normalized coordinates (x, y) of all 4 vertices.

To batch sequences of different lengths together, we pad sentences with the *end* token up to the same length. For example, given a batched sequences where the maximal length is 8, we pad the shorter sequence with 4 visible points only with two extra **E**:

S, P, P, P, P, E, E, E

We use the same Transformer decoder as the parallel models, but adopt the *masking* technique for parallel training, which prevents each token from attending to subsequent positions [30]. The token embedding is a fixed length vector of size 256 as shown in figure 3.

The auto-regressive approach predicts all polygons in a certain order. If not specified otherwise, we always predict objects going from left to right in the image, splitting ties with top-to-bottom order. We sort polygons and gates accordingly by their centers, computed as the average of their vertices. Particularly, for the polygon dataset, we also impose an order on **points** such that the polygon is clockwise defined by the points, otherwise the same set of the **points** may result in several polygons of different shapes.

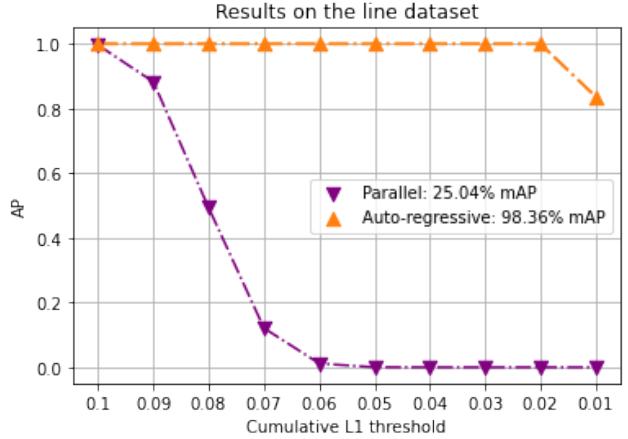


Figure 4: **Results on the line dataset.** The task is to predict a series of ordered points. The auto-regressive model produces approximately perfect prediction, while the parallel model fails.

5. Experiments

First, we validate that the auto-regressive approach is preferable on the line dataset where there is an explicit natural order in the predictions, and that the parallel solution is better on the point dataset for pure set prediction tasks without orders. We then compare the two strategies on more challenging gate datasets where the element is no longer a single point and where data is scarce. Finally, we explore deeper into the auto-regressive solution to study upon the importance of the prediction order and the order of the conditional variables.

On toy datasets, we train all models with 3,000,000 images. On the real gate datasets with 26,000 images, we train all models for 300 epochs. We apply multiple data augmentation techniques, including horizontal flip, vertical flip, hue shift, Gaussian noising.

We train all models on a single NVIDIA GeForce RTX 2080 Ti GPU with AdamW [22], and set the Transformer’s learning rate to 10^{-4} , the backbone’s learning rate to 10^{-5} , and the weight decay to 10^{-4} . The learning rate is dropped by a factor of 10 after 200 epochs, or after 2,000,000 images for the toy settings. Mask R-CNN [15] is also trained for 300 epochs with learning rate 5×10^{-3} , which is decreased by 10 after 200 epochs.

5.1. Evaluation

The evaluation metric on the gate and polygon datasets is mean Average Precision [4], averaged over different IoU thresholds ($[0.50, 0.55, \dots, 0.95]$). On the point and line datasets, we also evaluate mean Average Precision, but averaged over different point-to-point L_1 distance thresholds

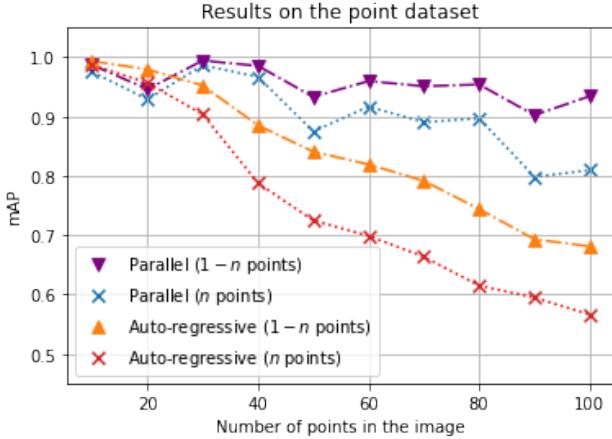


Figure 5: **Results on the point dataset.** The auto-regressive model shows marginally better performance than the parallel one on collections of limited cardinality, but its performance is lacking when the cardinality grows. The legends with *n points* represent results on images with exactly *n* points while the others indicate that the number of points in an image varies from 1 to *n*.

([0.10, 0.09, ..., 0.01]), computed directly on the relative coordinates.

5.2. Line and point detection

With this experiment we show that the conditional decomposition of collections by auto-regressive Transformers is beneficial when the elements in a collection adhere to a natural order. Results on the line dataset are shown in figure Fig. 4. For this task, a line prediction is considered as a false positive if the sum of all L_1 point distances is greater than the given threshold. The experiment shows that in this setting the auto-regressive solution is much more precise than the parallel counterpart as it is able to achieve perfect mAP up to a threshold of 0.02.

On the other hand, the experiments on the point dataset study the behaviour of parallel and auto-regressive models on a pure set prediction task. The models are now expected to predict the *n* points in any order. The parallel Transformers is order-insensitive because of the Hungarian matching, but the auto-regressive Transformers are trained by always feeding points in the left-to-right order. Results on this dataset are shown in Fig. 5. In this experiment, we train all models on images with 1 to *n* points. We present test results on images with 1 to *n* points, as well as results on images with exactly *n* points. Results show that on collection prediction tasks, the auto-regressive approach is effective when the cardinality is low, but quickly deteriorates as the cardinality increases.

The two experiments on lines and points prove that the

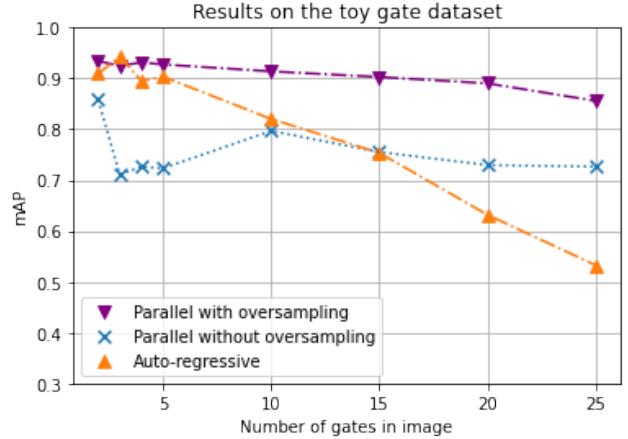


Figure 6: **Results on the gate dataset.** Each point of the plot is an experiment in which the model is trained and tested on images with 1 to *n* gates. They show the necessity of oversampling for parallel Transformers. The performance of auto-regressive models deteriorates when collection cardinality increases.

presence of a natural order in the task is indeed an important discriminative factor towards the performance of the parallel and auto-regressive models in collection prediction. Moreover, we show the advantage of the auto-regressive approach in low-cardinality prediction tasks. The parallel ones show significant advantage in predicting high-cardinality collections but at the cost of using redundant object queries (100 object queries in this experiment).

5.3. Gate detection

The following experiments verify whether the observations on the line and point datasets generalize in complicated scenarios where the collection element is a gate of four vertices. First, we evaluate the performances of auto-regressive approach and its parallel counterpart on the toy gate dataset, as shown in Fig. 6. There are two parallel models in comparison: one with oversampling where we use 30 object queries which is several times more than the total number of gates, and the other one without oversampling where we use as many object queries as the maximal number of objects in the images. The parallel model with oversampling outperforms the one without oversampling substantially, validating the need for the parallel models to oversample the cardinality. Moreover, we show once more the auto-regressive approach performs comparably to the parallel one on low-cardinality collections, but suffers from the growing cardinality.

In Fig. 7, we provide our findings on the generalization capabilities of these models, where the number of gates in test images is up to twice more than the number during

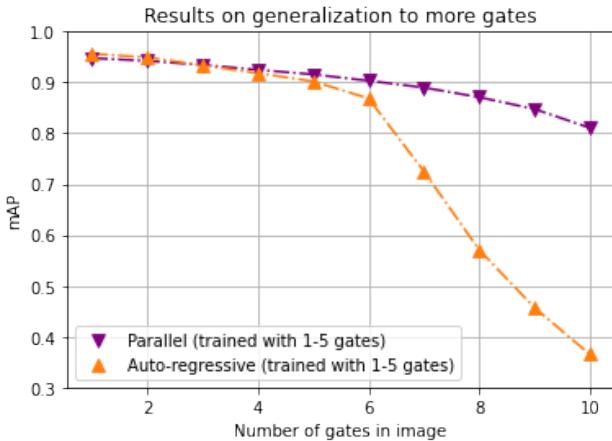


Figure 7: Generalization capability of the different models. We test both parallel and auto-regressive on images with twice as many gates as the training images. The parallel model is able to detect many more gates than the auto-regressive one. We believe oversampling attributes to better generalization.

training. The auto-regressive model struggles at detecting more gates, while the parallel one only shows minor performance decrease and achieves an mAP of 0.8 even on images with twice the amount of gates. We believe this behavior is a result of the oversampling strategy: each of the 30 object queries is assigned to at least one gate during training, which implicitly tells the model that there could be more objects than the ones it sees in each image.

Finally, we numerically compare the general performances of all models on both toy and synthetic gate datasets with images containing 1 to 6 gates. We choose Mask R-CNN as the baseline. Tab. 1 shows parallel Transformer outperforms Mask R-CNN by 5 % mAP on the toy dataset and by 4 % mAP on the synthetic gate dataset. The auto-regressive approach is not able to achieve comparable results as there is no natural order in this task. However, the auto-regressive model is approximately two times faster than the parallel model as it does not require Hungarian algorithm or oversampling.

5.4. Polygon detection

Results on the polygon toy setting dataset are shown in table Tab. 2. This experiment explores an hybrid scenario in which the collection elements we are trying to predict are also collections with an imposed order, because a collection of points can represent multiple polygons if the order of its points is not given. The auto-regressive approach is the special one described at the end of section 4.2, which predicts a polygon as a sequence of points followed by the *end-of-polygon* token, and it outperforms the parallel model

substantially by over 20 % in terms of mAP. When using the auto-regressive solution on polygons, we are predicting each polygon individually by predicting its vertices and then an end-of-polygon token. This means that we are splitting a problem of directly predicting all points of all polygons into multiple sub-problems of predicting each polygon and then deciding when all points have been predicted. We speculate that the advantage of the auto-regressive model could be a direct consequence of the conditional decomposition of the joint probability, as imposing the conditional order on these models can serve as a strong inductive bias by reducing the search space of the model. This behaviour might not show up on the previous experiments because gates, points and lines are simple collection elements which do not enlarge the search space enough. As polygon detection is representative of many common computer vision tasks, we leave further exploration of this approach as future work.

5.5. Orders of conditional decomposition

We conclude our experiments by providing insights on the importance of orders imposed on conditional decomposition of a collection by the auto-regressive Transformers. Moreover, we also study the influence of positional encodings. We run experiments on the toy gate dataset, synthetic gate and the polygon datasets multiple times, by using two different object orderings and by adding or removing the positional encodings of the attention layer. Without the positional encodings, the Transformer decoder is unaware of the index of the previous predictions in the sequence. When predicting a new point, it only knows which points were predicted before, but not their orders. The two orderings are the left-to-right, top-to-bottom order and the small-to-large order in terms of the size of the objects. Experiment results are shown in table Tab. 3. Using the left-to-right, top-to-bottom order and removing the positional encodings lead to the best performance in all experiments. We show the mean absolute difference of other models compared to best one. The result shows that the artificial order *matters* for auto-regressive models. Moreover, we show that fixed positional encodings are not beneficial in this setting, in contrast to the original Transformer architectures. This is expected as different orders of the same words in NLP can have different meaning, while this does not matter in our setting, as knowing the location of a point can already prevent our model to predict the same point twice.

6. Conclusion

We studied two important variants of the Transformer architecture: the parallel one and the auto-regressive one on the polygonal shape prediction task cast as a collection prediction problem. We find that the ordering of objects is a strong factor towards the performance of these models on point, line, gate and polygon datasets. The auto-regressive

Models	Toy gates	Training time [min/10k images]	Synthetic gates	Training time [min/10k images]
<i>Mask R-CNN</i>	90.67%	23.6	61.30%	13.85
Parallel	95.45%	2.2	65.00%	3.85
Auto-regressive	87.67%	1.5	59.03%	1.85

Table 1: **Numerical results on gate datasets.** We evaluate the overall performance in mAP averaged over 10 IoU thresholds $[0.50, 0.55, \dots, 0.95]$. The parallel model outperforms Mask R-CNN. The performance of auto-regressive model is lacking due to the absence of natural order in this task.

Models	Polygons dataset	Training time [min/10k images]
Parallel	53.55%	5.1
Auto-regressive	76.41%	2.15

Table 2: **Numerical results on the polygon dataset.** Scores are represented as mAP averaged over 10 IoU thresholds $[0.50, 0.55, \dots, 0.95]$. The auto-regressive approach outperforms the parallel one by over 20% absolute mAP. We speculate that this could be a direct consequence of conditional decomposition of the collection.

	With positional encodings	W/o positional encodings
Order based on polygon positions	$-1.90\% \pm 1.07\%$	—
Order based on the size of polygons	$-7.68\% \pm 4.53\%$	$-2.95\% \pm 1.30\%$

Table 3: **Ablation studies on position encodings and condition orders.** This tables shows results of adding positional encodings and imposing orders. The top performing model is spatially ordered (left-to-right and top-to-bottom) and has no positional encoding. Our observations are that: (1) changing the imposed order on auto-regressive Transformers has a great impact on the performance as shown in the first column; (2) adding position encodings decreases the performance.

model benefits collections of small cardinality or on collection prediction problems that can be easily split into multiple easier tasks.

One limitation of this research is that most experiments are conducted on toy datasets only, and it is unclear whether the observations and conclusions on toy datasets generalize on challenging real-world datasets. A task that would be suitable to expand our research is polygonal building segmentation from satellite images, as in [11, 23].

As future work, we find it important to further explore on the polygon detection task as it is a fundamental problem that is representative in computer vision. Testing the effect of fixed or learned token embeddings would also be of interest.

References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [3] Zhongzheng Cao, Rui Wang, Xiangyang Wang, Zhi Liu, and Xiaoqiang Zhu. Improving human pose estimation with self-attention generative adversarial networks. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 567–572. IEEE, 2019.
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [5] William Chan, Chitwan Saharia, Geoffrey Hinton, Mohammad Norouzi, and Navdeep Jaitly. Imputer: Sequence modelling via imputation and dynamic programming. In *International Conference on Machine Learning*, pages 1403–1413. PMLR, 2020.
- [6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Philipp Dürnay. Detecting empty wireframe objects on micro-air vehicles: Applied for gate detection in autonomous drone racing. 2018.

[10] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

[11] Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building segmentation by frame field learning. *arXiv preprint arXiv:2004.14875*, 2020.

[12] Nicolas Girard and Yuliya Tarabalka. End-to-end learning of polygons for remote sensing image classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2083–2086. IEEE, 2018.

[13] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[14] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

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

[16] Vladimir Iglovikov, Selim Seferbekov, Alexander Buslaev, and Alexey Shvets. Ternausnetv2: Fully convolutional network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 233–237, 2018.

[17] Sungwoo Jung, Sungwook Cho, Dasol Lee, Hanseob Lee, and David Hyunchul Shim. A direct visual servoing-based framework for the 2016 iros autonomous drone racing challenge. *Journal of Field Robotics*, 35(1):146–166, 2018.

[18] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005.

[19] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[20] Zuoyue Li, Jan Dirk Wegner, and Aurélien Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1715–1724, 2019.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[23] Sharada Prasanna Mohanty, Jakub Czakon, Kamil A Kaczmarek, Andrzej Pyskir, Piotr Tarasiewicz, Saket Kunwar, Janick Rohrbach, Dave Luo, Manjunath Prasad, Sascha Fleer, et al. Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence*, 3, 2020.

[24] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[26] S Hamid Rezatofighi, Vijay Kumar BG, Anton Milan, Ehsan Abbasnejad, Anthony Dick, and Ian Reid. Deepsetnet: Predicting sets with deep neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5257–5266. IEEE, 2017.

[27] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[28] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[29] Wei Tang and Ying Wu. Does learning specific features for related parts help human pose estimation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1107–1116, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[31] Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pages 6487–6494. PMLR, 2019.

[32] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

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

[34] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. *arXiv preprint arXiv:1703.06114*, 2017.

[35] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Deep set prediction networks. *arXiv preprint arXiv:1906.06565*, 2019.

[36] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Fspool: Learning set representations with featurewise sort pooling. *arXiv preprint arXiv:1906.02795*, 2019.

[37] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.