DiffuDETR: Rethinking Detection Transformers with Denoising Diffusion Process

Youssef Nawar, Mohamed Badran, Marwan Torki

Agenda

- Introduction
- Related Work
- Approach
- Results
- Conclusion

Agenda

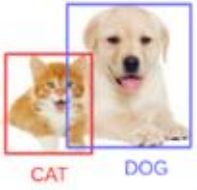
- **Introduction**
- Related Works
- Approach
- Results
- Conclusion

Introduction

Object detection is a fundamental task in computer vision, it is essential to understand the current state of object detection and why a new approach is necessary.

- **The Goal**: Object detection requires a model to perform two simultaneous tasks: **Classification** (What is this?) and **Localization.**

Image Localization



Object Detection

Introduction

- Older Methods: Depend on predefined anchor boxes, selective search, and NMS post-processing limiting flexibility and generalization.
- DETR End-to-end set prediction with bipartite matching, removes the need of anchors or NMS. But queries are zero-initialized, causing slow convergence.

Limitations

- Query Initialization DETR variants must learn query-to-object alignment from scratch. Spatial priors in queries dramatically improve convergence.

- DiffuDETR: Model query initialization as a diffusion denoising process. Generate reference points from Gaussian noise, guided by image features.

Agenda

- Introduction
- **Related Works**
- Approach
- Results
- Conclusion

DETR

- **DETR Evolution**: Moves from bipartite set matching to dynamic anchor boxes (DAB-DETR) and auxiliary denoising tasks (DN-DETR, DINO)
- While earlier work refined noisy proposal boxes, we adapt diffusion specifically to the **object queries** of transformer-based detection

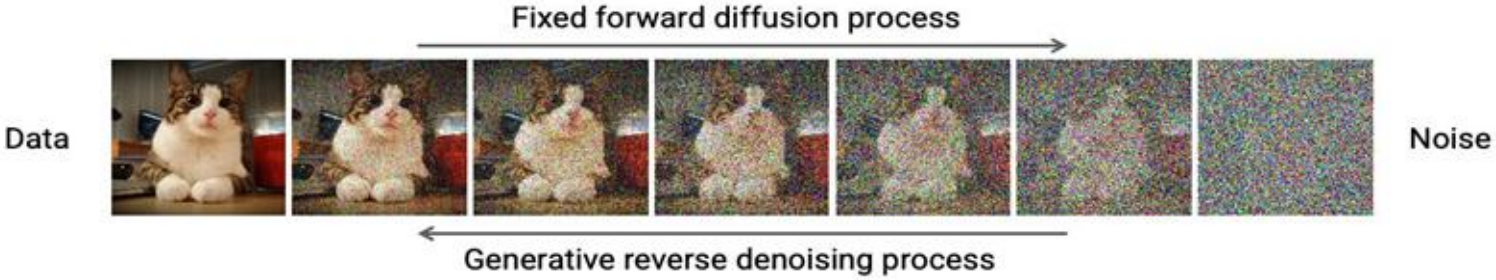
Generative Models for Object Detection

- Sequence-to-Sequence (Pix2Seq): Formulated detection as a language modeling task, predicting bounding box coordinates as discrete tokens.

Pix2Seq

Denoising Diffusion Probabilistic Models (DDPM)

- **The Foundation:** Diffusion models (like those in Stable Diffusion) learn to reverse a Gaussian noise process to generate high-fidelity data.



Denoising Diffusion Probabilistic Models (DDPM)

- Diffusion-Based Detection (DiffusionDet): The first to treat detection as a generative process. How it works: Refines noisy bounding boxes directly in the image space using a convolutional or hierarchical backbone.
- DiffuDETR operates on the latent object queries and reference points within the Transformer architecture.

Agenda

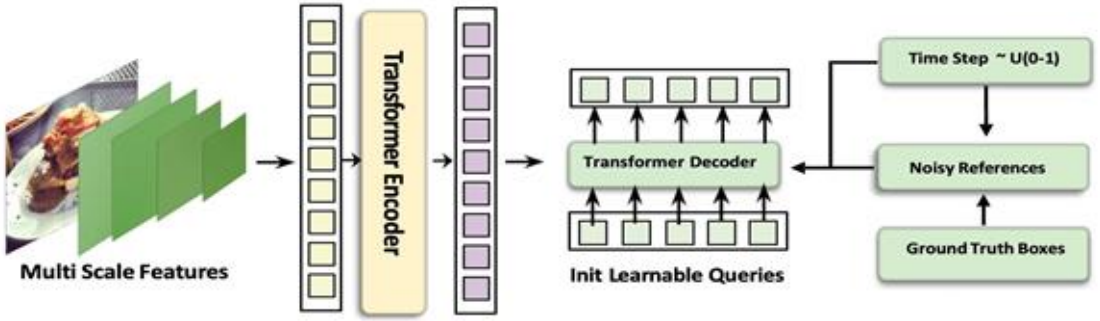
- Introduction
- Related Works
- **Approach**
- Results
- Conclusion

Approach

The training strategy in **DiffuDETR** transforms object detection into a generative task by teaching the model to recover precise spatial anchors from corrupted data.

- **Forward Diffusion (The Noise Step)**:
  - We treat the ground-truth (GT) bounding box centers as the "clean" data distribution.
  - Controlled levels of gaussian noise are added to these GT coordinates based on a predefined variance schedule (t).
  - This generates a set of **noisy reference points** that serve as the initial spatial priors for the decoder.
- **Conditional Query Generation**:
  - The Transformer decoder is conditioned on two signals: the **visual features** (from the image backbone) and the **noisy reference points**.
  - A **Time-step Embedding** is injected into the queries, allowing the decoder to "know" the current noise level and adjust its denoising strength accordingly.

- **The Denoising Objective**:
  - The model is optimized to predict the original, clean coordinates from the noisy versions.
  - Unlike standard DETR which uses a fixed set of learnable queries, our queries are dynamically tied to the diffusion time-step, leading to a more robust representations.

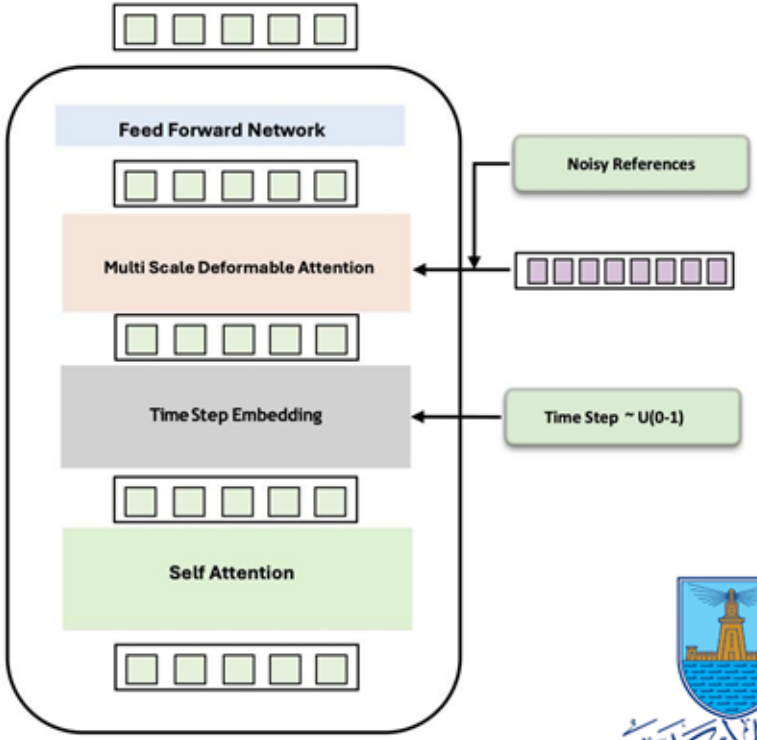


Inference

Inference in **DiffuDETR** balances the generative power of diffusion with the real-time requirements of object detection. We move from random noise to precise object queries using a deterministic and optimized pipeline.

- **From Noise to Predictions**:
  - The process begins by sampling a set of reference points from a **Gaussian distribution** N(0,I).
  - These noisy points are fed into the Transformer decoder along with the image features.
- **Deterministic DDIM Sampling**:
  - Instead of stochastic sampling, we employ the **Denoising Diffusion Implicit Model (DDIM)**.
  - This allows for a deterministic mapping from noise to object locations, ensuring stable and reproducible detections.

- **The Lightweight Sampling Scheme**:
  - To maintain high FPS (frames per second), we avoid the hundreds of steps typically used in image generation.
  - We introduce a scheme that requires only **multiple forward passes through the decoder** drastically reducing computational overhead.
- **Zero-Cost Accuracy Gains**:
  - Even with a single step (T=1), the diffusion-trained weights provide a better spatial prior than standard learnable queries.
  - Increasing to a small number of iterative steps allows the model to "self-correct" and refine box boundaries for complex scenes without re-processing the image through the backbone.



جامعة الإسكندرية
ALEXANDRIA UNIVERSITY

Agenda

- Introduction
- Related Works
- Approach
- **Results**
- Conclusion

Datasets

To validate **DiffuDETR's** robustness and generalization capabilities, we evaluated the framework across three distinct and challenging large-scale datasets. Our approach consistently outperformed baselines, particularly in scenarios with high category density and object occlusion

COCO 2017 (Common Objects in Context)

- **Focus:** Standard benchmark for general object detection.
- 80 classes · 118K train · 5K val

جامعة الإسكندرية
ALEXANDRIA UNIVERSITY

Datasets

LVIS v1.0 (Large Vocabulary Instance Segmentation)

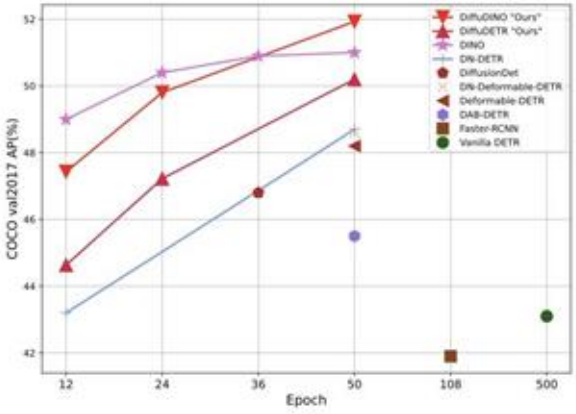
- **Focus:** Challenges models with a "long-tail" distribution of **1,203 object categories**.
- 1,203 classes · 100K train · 20K val

V3Det

- **Focus:** A vast dataset featuring **13,204 categories**, testing the extreme limits of openworld and large-vocabulary detection.
- 13,204 classes · 183K train · 30K val

COCO2017

Model	Epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ResNet-50 (He et al., 2016)							
DETR-DC5 (Carion et al., 2020)	500	43.3	63.1	45.9	22.5	47.3	61.1
DN-Deformable DETR (Li et al., 2022)	50	48.6	67.4	52.7	31.0	52.0	63.7
MS-DETR (Zhao et al., 2024)	24	50.9	68.4	56.1	34.7	54.3	65.1
Salience DETR(Hou et al., 2024)	24	51.2	68.9	55.7	33.9	55.5	65.6
MR-DETR (Zhang et al., 2025)	24	51.4	69.0	56.2	34.9	54.8	66.0
Pix2Seq (Chen et al., 2021)	300	43.2	61.0	46.1	26.6	47.0	58.6
DiffusionDet (Chen et al., 2023b)	-	46.8	65.3	51.8	29.6	49.3	62.2
Deformable DETR (Zhu et al., 2020)	50	48.2	67.0	52.2	30.7	51.4	63.0
Align-DETR (Cai et al., 2024)	24	51.4	69.1	55.8	35.5	54.6	65.7
DINO (Zhang et al., 2022)	36	50.9	69.0	55.3	34.6	54.1	64.6
DiffuDETR (Ours)	50	50.2	66.8	55.2	33.3	53.9	65.8
DiffuAlignDETR (Ours)	24	**51.9**	69.2	**56.4**	34.9	55.6	66.2
DiffuDINO (Ours)	50	**51.9**	**69.4**	55.7	**35.8**	**55.7**	**67.1**



ResNet-101 (He et al., 2016)							
DETR-DC5 (Carion et al., 2020)	50	43.5	63.8	46.4	21.9	48.0	61.8
DAB-DETR-DC5 (Liu et al., 2022)	50	46.6	67.0	50.2	28.1	50.5	64.1
DN-DETR-DC5 (Li et al., 2022)	50	47.3	67.5	50.8	28.6	51.5	65.0
MR-DETR (Zhang et al., 2025)	12	51.4	68.6	55.7	**34.3**	55.1	66.7
Pix2Seq (Chen et al., 2021)	300	44.5	62.8	47.5	26.0	48.2	60.3
DiffusionDet (Chen et al., 2023b)	-	47.5	65.7	52.0	30.8	50.4	63.1
DINO (Zhang et al., 2022)	12	50.0	67.7	54.4	32.2	53.4	64.3
Align-DETR (Cai et al., 2024)	12	51.2	68.8	55.7	32.9	55.1	66.6
DiffuDINO (Ours)	12	51.2	68.6	55.8	33.2	**55.6**	**67.2**
DiffuAlignDETR (Ours)	12	**51.7**	**69.3**	**56.1**	34.0	**55.6**	67.0

LVIS

Model	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP_r	AP_c	AP_f
ResNet-50 (He et al., 2016)									
DINO (Zhang et al., 2022)	26.5	35.9	27.8	20.0	35.2	40.9	9.2	24.6	36.2
DiffuDINO (Ours)	**28.9**	**38.5**	**30.8**	**20.7**	**37.5**	**46.4**	**13.7**	**27.6**	**36.9**
ResNet-101 (He et al., 2016)									
DINO (Zhang et al., 2022)	30.9	40.4	32.8	23.2	40.5	46.3	**13.9**	29.7	39.7
DiffuDINO (Ours)	**32.5**	**42.4**	**34.8**	**23.5**	**43.4**	**49.7**	13.5	**32.0**	**41.5**

V3 Det

Model	AP	AP_{50}	AP_{75}
ResNet-50 (He et al., 2016)			
DINO (Zhang et al., 2022)	33.5	37.7	35.0
DiffuDINO (Ours)	**35.7**	**41.4**	**37.7**
Swin-B (Liu et al., 2021)			
DINO (Zhang et al., 2022)	42.0	46.8	43.9
DiffuDINO (Ours)	**50.3**	**56.6**	**52.9**

جامعة الإسكندرية
ALEXANDRIA
UNIVERSITY

Agenda

- Introduction
- Related Works
- Approach
- Results
- **Conclusion**

Conclusion

We reformulate object detection as object generation problem which eliminates any needs to learn any spatial prior. DiffuDETR recasts DETR query initialization as denoising diffusion a fundamentally new perspective on a long-standing problem.

Future Work

- Integrate generative / autoregressive approaches
- Explore instance segmentation
- Faster Convergence

Thank you for your attention

Feel free to reach me on youssof.nawar@tum.de