

# TIED: A Cycle Consistent Encoder-Decoder Model for Text-to-Image Retrieval

Clint Sebastian<sup>1,2</sup>, Raffaele Imbriaco<sup>1</sup>, Panagiotis Meletis<sup>1</sup>,  
Gijs Dubbelman<sup>1</sup>, Egor Bondarev<sup>1</sup>, Peter H.N. de With<sup>1,2</sup>  
<sup>1</sup>VCA Group, Eindhoven University of Technology  
<sup>2</sup>Cyclomedia B.V

F.c.sebastian, r.imbriaco, p.c.meletis@tue.nl

## Abstract

*Retrieving specific vehicle tracks by Natural Language (NL)-based descriptions is a convenient way to monitor vehicle movement patterns and traffic-related events. NL-based image retrieval has several applications in smart cities, traffic control, etc. In this work, we propose **TIED**, a **text-to-image encoder-decoder model** for the simultaneous extraction of visual and textual information for vehicle track retrieval. The model consists of an encoder network that enforces the two modalities into a common latent space and a decoder network that performs an inverse mapping to the text descriptions. The method exploits visual semantic attributes of a target vehicle along with a **cycle-consistency loss**. The proposed method employs both intra-modal and inter-modal relationships to improve retrieval performance. Our system yields competitive performance achieving the 7th position in the Natural Language-Based Vehicle Retrieval public track of the 2021 NVIDIA AI City Challenge. We demonstrate that the proposed TIED model obtains six times higher Mean Reciprocal Rank (MRR) than the baseline, achieving an MRR of 15.48. The code and models will be made publicly available.*

## 1. Introduction

Vehicle track retrieval from traffic cameras [9] is an essential component of upstream systems aiming for urban planning and traffic-flow control. Large-scale retrieval of vehicle tracks is difficult to obtain with conventional image or video retrieval methods, due to the immense variety of motion patterns and vehicle semantics that need to be considered. Descriptions for these tracks in Natural Language (NL) is an appealing alternative method to enable the retrieval system to directly interact with human-given descriptions [33, 2]. The objective of **NL-based vehicle track retrieval** [9] is to match a given NL description to the corre-

Figure 1: Qualitative vehicle retrieval results using the baseline method (Rows 1, 3) and our method (Rows 2, 4) of a frame from the retrieved tracks. The queries are “A blue sedan runs down the street.” and “A red cargo truck pulls a yellow cement mixer.”

sponding vehicle track. The NL description is given as one or more text queries, and the vehicle tracks are a sequence of frames from a single camera, where the location of the vehicle is known. This task combines visual and textual modalities, thus solutions should simultaneously account for intra- and inter-modality challenges. Vehicle tracks include a wide variety of vehicle types, colors, and motion types. NL queries often have variations and ambiguities, since different people can describe the same vehicle seman-

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

~~~~~

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_









