Global Dilated Attention and Target Focusing Network for Robust Tracking

Yun Liang*, **Qiaoqiao Li**, **Fumian Long**

Guangzhou Key Laboratory of Intelligent Agriculture, College of Mathematics
and Informatics, South China Agricultural University
yliang@scau.edu.cn, qiaoqiaoli23@163.com, lfm_sunny@stu.scau.edu.cn

Abstract

Self Attention has shown the excellent performance in tracking due to its global modeling capability. However, it brings two challenges: First, its global receptive field has less attention on local structure and inter-channel associations, which limits the semantics to distinguish objects and backgrounds; Second, its feature fusion with linear process cannot avoid the interference of non-target semantic objects. To solve the above issues, this paper proposes a robust tracking method named GdaTFT by defining the Global Dilated Attention (GDA) and Target Focusing Network (TFN). The GDA provides a new global semantics modeling approach to enhance the semantic objects while eliminating the background. It is defined via the local focusing module, dilated attention and channel adaption module. Thus, it promotes semantics by focusing local key information, building long-range dependencies and enhancing the semantics of channels. Subsequently, to distinguish the target and non-target objects both with rich semantics, the TFN is proposed to accurately focus the target region. Different from the present feature fusion, it uses the template as the query to build a point-to-point correlation between the template and search region, and finally achieves part-level augmentation of target feature in the search region. Thus, the TFN efficiently augments the target embedding while weakening the non-target objects. Experiments on challenging benchmarks (LaSOT, TrackingNet, GOT-10k, OTB-100) demonstrate that the GdaTFT outperforms many state-of-the-art trackers and achieves leading performance. Code will be available.

1 Introduction

Visual tracking is a fundamental task in computer vision, aiming to predict the state of a target in video sequences given its initial state. It has been widely used in various applications such as visual surveillance and autonomous driving. Many efforts have been done in recent years, however, developing a robust and accurate tracker is still challenging due to the various hinders such as rapid deformations, occlusions and background clutters that often occur in tracking.

Recently, due to the good balance between accuracy and speed, siamese network based trackers have drawn great attention. These methods (Bertinetto et al. 2016; Li et al. 2018; Chen et al. 2020; Guo et al. 2020) map the template and the search region to the same feature space, and achieve semantic-level feature matching through cross correlation to locate the target. Therefore, it is critical to extract robust semantic information in tracking. However, limited by fixed operation mode of convolution and small receptive field, these methods are difficult to extract semantic information effectively. Thanks to the flexible adaptability comparing to fixed weights of convolution, attention mechanism can obtain the dynamic weights according to different inputs, which helps to extract robust semantics. By introducing attention mechanism, the methods (Wang et al. 2018; Choi et al. 2017; Fu et al. 2021; Du et al. 2020) adaptively enhance feature with semantic information. But the challenge raised by small receptive field remains unresolved, which still limits the semantics extraction of feature. Therefore, due to its excellent ability to build long-range dependencies, self attention has been widely used in visual tracking. It is introduced to perform global semantics modeling and achieve feature fusion between template and search region to replace cross correlation, which yields great performance (Yan et al. 2021; Wang et al. 2021; Yu et al. 2021). However,

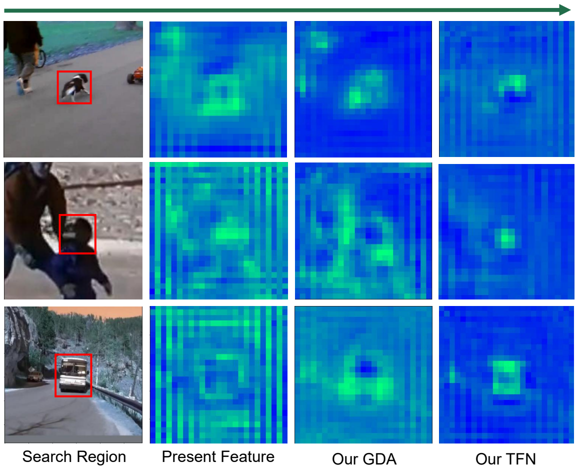


Figure 1: Features of our GdaTFT. Compared with present feature, our GDA distinguishes objects and background by enhancing semantics. Then, our TFN further augments target embedding by weakening no-target semantic objects.

*Corresponding Author: Yun Liang.

even with its great success, those self-attention-based methods still suffer from the following two typical problems:

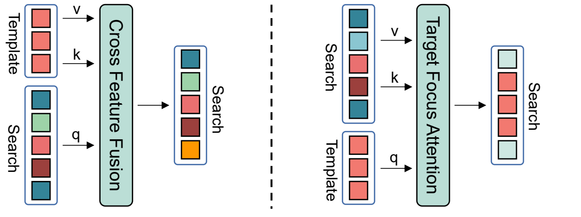
(1) Their global receptive field (Yu et al. 2020; Cui et al. 2021) brings less attention on local structural information and ignores inter-channel associations, which limits the semantics that can effectively distinguish objects and backgrounds. (2) These methods (Guo et al. 2021; Wang et al. 2021) take the feature of search region as query to perform linear-process on the template feature, and use feature fusion to enhance the target embedding. This process is difficult to enhance target embedding effectively and easy to introduce noise from search region.These problems make features cannot focus on target as the second column in Fig. 1.

To address the above issues, this paper proposes a robust tracking method named GdaTFT by defining the Global Dilated Attention (GDA) and Target Focusing Network (TFN). The GDA is proposed to achieve comprehensive feature semantics extraction from both space and channel dimensions. With the proposed dilated attention, it performs dilated sampling on feature map to split it into different groups, and then builds the global and sparse long-range dependencies within these independent groups to enhance the semantics of all objects. In addition, the GDA builds the inter-channel associations to adaptively select and enhance channels with rich semantics via our channel adaption module.

Furthermore, the TFN is defined to enhance the target embedding on the semantics-enhanced search region feature which is computed by the GDA. As Fig. 2, it takes template as the query to transfer the target information from the template to the search region and thereby obtains a point-to-point similarity between them, which helps to weaken the expression of non-target objects in the search region. Comparing with feature fusion based method, the TFN can effectively avoid interference from other semantic objects.

Our main contributions can be summarized as follows.

- We define the Global Dilated Attention (GDA) to provide a global semantics modeling to enhance the semantics of feature. It greatly enhances the expression of semantic objects and eliminates the interference of non-semantic background to improve the robustness of our tracker.
- We define the Target Focusing Network (TFN) to construct point-to-point associations between template and search region. It effectively distinguishes target and non-target objects both with rich semantics, and further successfully achieves part-level target embedding augmentation based on the semantics-enhanced feature from GDA.
- A novel tracking method GdaTFT is proposed via the GDA and TFN. Experiments on challenging benchmarks (LaSOT, TrackingNet, GOT-10k, OTB-100) demonstrate that our GdaTFT outperforms most of the state-of-the-art trackers and achieves leading performance.



(a) Feature Fusion (b) Target Focusing Network (Ours)

Figure 2: Model comparison between (a) feature fusion based method and (b) our TFN. Compared with (a), the TFN produces more accurate search region with less noises and distractors (such as the cyan and yellow part in "Search") by using the feature of template as the query.

2 Related Work

Visual Tracking based on Siamese Network. In recent years, the siamese network has attracted most of the attention in visual tracking due to its excellent performance. As the pioneer, SiamFC (Bertinetto et al. 2016) constructs a simple siamese network based on cross correlation. Firstly, the patches of template and search region are input into a siamese backbone network to extract feature, and then the feature of template is used as convolution kernel to perform convolution operation on the feature of search region to obtain the response map, which is used to predict the target position. The above is the main framework of the siamese-based trackers. Many works have been extended to make great progress based on the siamese tracking framework. Because of the outstanding performance in dealing with object scale changes, Region Proposal Network (RPN) (Ren et al. 2015) is widely introduced to improve visual tracking recently. SiamRPN (Li et al. 2018) combines the RPN with siamese network, and uses depth-wise cross correlation for feature fusion, which improves the tracking accuracy. Based on SiamRPN, SiamRPN++ (Li et al. 2019) and SiamDW-RPN (Li et al. 2019) both explore tracking by deeper network to improve tracking performance. However, RPN-based trackers often limited by anchor-related hyper-parameters that require complex manual tuning. Furthermore, many approaches have turned their attention to anchor-free designs, by treating the task of tracking as an integrated problem of classification and regression. SiamFC++ (Xu et al. 2020), SiamBAN (Chen et al. 2020) and SiamCAR (Chen et al. 2020) are all focused here, and the anchor-free network achieves the state-of-the-art performance at that time. In this paper, we adopt the anchor-free siamese-based tracking architecture, and avoid complex tuning of hyper-parameters.

Attention Mechanism and Visual Tracking. Recently, attention mechanism has been widely used in visual tracking. Compared with convolution operation, attention mechanism owns excellent adaptive ability to flexibly focus on the important things according to the input. CGACD (Du et al. 2020) applies attention on the results of cross correlation and uses its output to enhance the search region feature. STMTrack (Fu et al. 2021)uses the attention to select the template that best matches the current frame from memory. SiamGAT (Guo et al. 2021) introduces graph attention into Siamese Networks to improve tracking results. However, the receptive field of traditional attention is limited, which limits the semantic extraction of these trackers. Therefore, the self attention with the ability to model global depen-

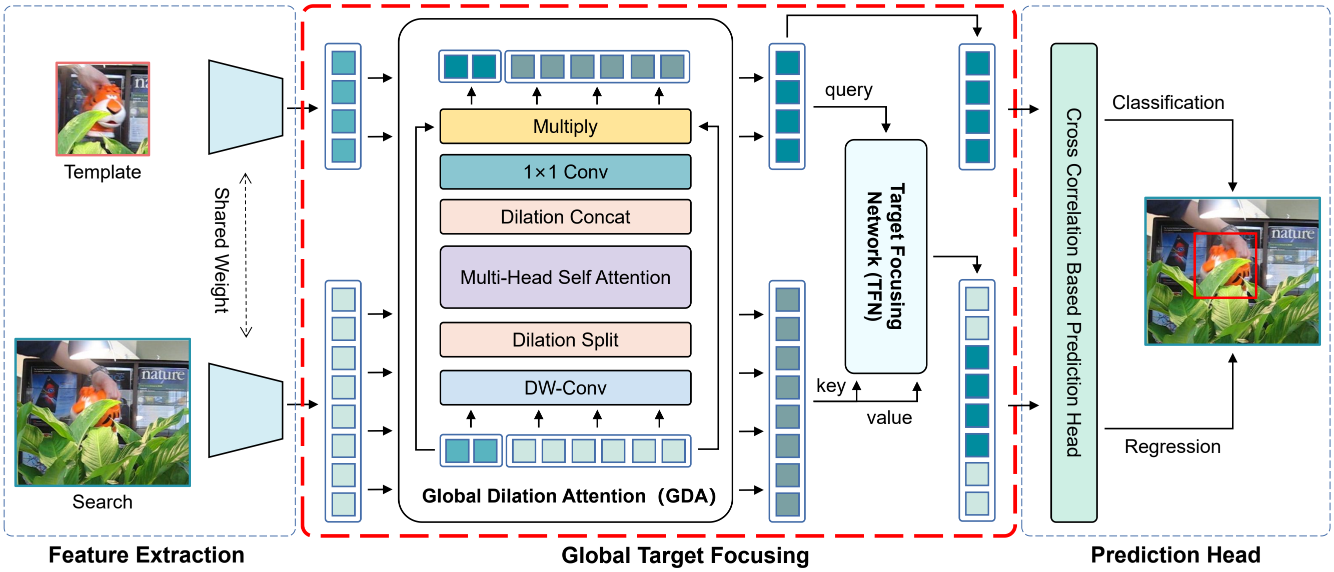


Figure 3: Architecture of our GdaTFT framwork. This framework contains three main components: siamese backbone network for feature extraction, global target focusing for global semantic enhancement and target focusing, and prediction head for target localization. The Global Target Focusing is the main work and constructed by our GDA and TFN (in the red rectangle above).

dencies has become very popular in visual tracking. TransT (Chen et al. 2021) designs a feature fusion network based on self attention to transfer information between search region and template, and enhances the tracking accuracy. Inspired by DETR (Carion et al. 2020), STARK (Yan et al. 2021) proposes a self-attention-based encoder-decoder network to achieve feature fusion between template and search region. With the encoder-decoder framework, TrDimp (Wang et al. 2021) introduces temporal cues and then uses self attention to achieve spatiotemporal information fusion at the same time. However, the current methods based on self attention also have two obvious problems. First, the self attention aims to build long-range dependencies, ignoring the local structural information and inter-channel associations, which will directly affect the extraction of feature semantic information. Second, the feature fusion based on self attention often easily introduces noises that lead to tracking errors. To address the above issues, this paper proposes a robust tracking network based on global dilated attention and target focusing network, and gradually implements the semantic modeling of feature and target focusing to achieve robust tracking.

3 Method

This section details the proposed tracker GdaTFT. As Fig. 3, it includes: siamese backbone for feature extraction, global target focusing module (composed of the proposed GDA and TFN) , and prediction head for target localization.

3.1 Overview

Our GdaTFT implements visual tracking by four steps. First, it extracts the feature of the template and search region separately using the siamese backbone network. Second, The GDA performs efficient feature augmentation for both template and search region feature to enhance the semantics of feature in both spatial and channel dimensions. Third, the TFN uses the template as a query to build a point-to-point correlation between the template and the search region, and transfers the information from the template to the search region to further enhance the feature of target. Finally, we input the output of TFN into the cross correlation based prediction head to produce the target region.

Feature Extraction. Like most siamese-based trackers, the proposed GdaTFT constructs a feature extraction network with two parameter-sharing branches, including a template branch and a search region branch, as in Fig. 3. The template branch takes the target patch of the initial frame as input, while the search region branch takes the search region of current frame as input. After training offline, this feature extraction network maps the inputs of the two branches to the same feature space to produce the input of GDA.

Global Target Focusing. This part is designed to achieve global semantics modeling and target focusing, as the red rectangle in Fig. 3, it is defined by the proposed GDA and TFN. In details, we input the feature computed from the two branches to GDA to enhance the semantics to distinguish the objects from background. Subsequently, using the output of GDA, our TFN enhances the target feature in the search region to further distinguish target object and non-target objects both with semantics. The relationship of GDA and TFN is progressive and complementary. They successively exclude the interference of background and other semantic objects, and furthermore, provide robust feature for subsequent prediction head.

Prediction Head. After being enhanced by GDA and TFN, the feature of non-target objects are greatly weakened, and the influence of background and distractors in the search region is reduced, which is of great significance for feature matching based on cross correlation. With the output feature from global target focusing, we employ a cross correlation

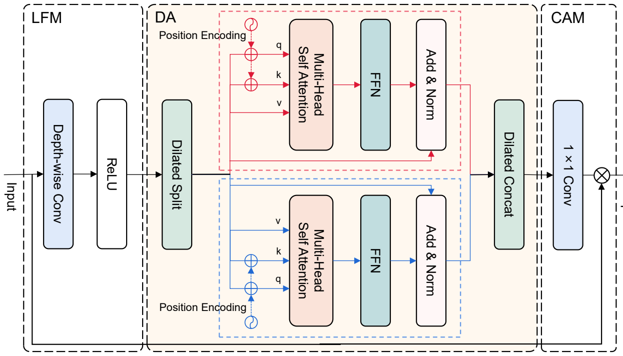


Figure 4: Architecture of GDA. It consists of three parts from left to right: local focusing module (LFM), Dilated Attention (DA), and channel adaptation module (CAM). The three jointly implement our global semantics modeling.

operation commonly used in siamese based trackers to generate a response map of target and search region. The region with the largest response in the response map is considered to be the most likely target. Finally, the response map is input into a simple prediction head composed of a classification branch and a regression branch to locate the target, so as to obtain the accurate target region.

3.2 Global Dilated Attention (GDA)

In the feature extraction phrase as Fig. 3, The proposed GdaTFT firstly uses the traditional siamese convolution neural network (CNN) to extract feature, which has a good performance in building local structures. However, due to the small receptive field, the CNN is difficult to build long-range dependencies, which limits the semantics of feature and the robustness of the tracker. To solve this problem, we propose GDA to optimize the features. It achieves global semantics enhancement of feature extracted by CNN both in space and channel dimensions, and improves the robustness of feature. As Fig. 4, the GDA consists of local focusing module, dilated attention and channel adaption module, which are used to gradually implement the construction of local structure and long-range dependencies and improve the inter-channel associations. We use GDA to enhance the semantics by:

$$F_{output} = F_{input} \cdot C\left(D\left(L\left(F_{input}\right)\right)\right) \tag{1}$$

where F_{input} and F_{output} are the input and output of GDA, L is the local focusing module, D is the dilated attention, and C is the channel adaptation module. The relationships of L, D and C are complementary, where L and D are used to implement global modeling in space, and C builds the correlations of feature between channels.

Local Focusing Module (LFM). We define the local focusing module by using depth wise convolution (DW-Conv) to build local structural information and local dependencies of feature. Compared with the traditional convolution where each channel corresponds to c convolution kernels, the channels and convolutions of DW-Conv have a one-to-one correspondence, so the parameter amount is only $1/c$ of the former. Assuming that the size of the convolution kernel is k, for each point on the feature map, LFM builds a connection between k^2 points centered on the point, enhances the local focusing ability of the network. It preserves the valuable local details for the feature semantics extraction.

Dilated Attention (DA). We propose dilated attention to build long-range dependencies of feature. Compared with local structural information, which focuses on object details, long-range dependencies pays more attention to the establishing of feature semantics. Feature semantics has stronger anti-interference ability than object details.

We construct the DA by the following three steps as Fig. 4. First, with dilated split, DA samples the feature map at intervals of d and splits it into d^2 independent groups. d is the key parameter of DA which represents the dilation rate of the module. Second, for group X_i, a multi-head self attention module (MSAM), which consists of a multi-head self attention (MSA), feed-forward network (FFN) and add & norm, is used to construct its global dependencies. It first performs a linear process on X_i as Equation 2 to get X_i^q, X_i^k and X_i^v, and inputs them to the MSA to construct dependencies as Equation 3. Subsequently, the result is input into the FFN and add & norm to get the output. The red and blue rectangles in Fig. 4 are two independent MSAMs and there are d^2 independent MSAMs in this part.

$$X_i^q = X_i^k = X_i + pos(X_i), X_i^v = X_i \tag{2}$$

$$\mathrm{MSA}(X_i) = softmax\left(\frac{X_i^q (X_i^k)^T}{\sqrt{d_x}}\right) X_i^v \tag{3}$$

where pos is the spatial position encoder for learning the position encoding information of X_i and d_x is the dimensionality of X_i. The MSAM is calculated by:

$$\mathrm{MSAM}(X_i) = X_i + \mathrm{FFN}(\mathrm{MSA}(X_i)) \tag{4}$$

Third, dilated concat is performed to put the sampled points in each group back to their corresponding positions in the feature map. The DA constructs sparse long-range dependencies of the feature by dilated sampling.

Compared with dense self attention, sparse feature dependencies focus more on feature semantics and avoid the interference of non-semantic noises from background. As the last column of Fig. 5, the augmented feature map by DA has clear boundaries, where the yellow, red and purple regions are classified as semantics, while the cyan regions are considered to non-semantics background and been eliminated.

Channel Adaptation Module (CAM). The LFM and DA together complete the global semantics modeling in the spatial dimension, but ignore the semantic differences between channels, which makes it difficult to extract the semantics of feature completely. Therefore, we design a channel adaptation module to construct inter-channel associations. Specifically, we introduce a 1×1 convolution to achieve information exchange between channels, and adaptively strengthen the feature expression of channels with semantics.

Totally, the proposed GDA implements global semantics modeling of feature in both spatial and channel dimensions. Compared with the current popular self attention, it builds sparse long-range dependencies within feature, and further constructs local structural information as well as

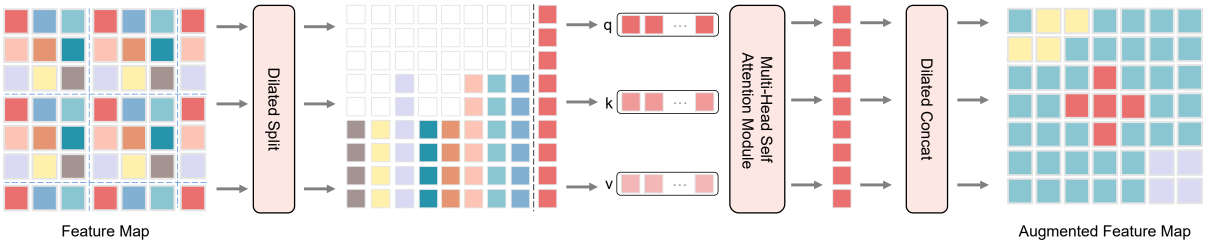


Figure 5: The proposed DA. First, assumed $d = 3$, the 7×7 feature map is divided into nine independent groups with different colors. Second, each group is linearly transformed to generate the corresponding q, k, v and sent to the multi-head self attention module for intra-group feature enhancement. Finally, all points of each group are put back to their original positions via dilated concatenation. As the last column, the yellow, red and purple regions describing semantic objects are enhanced, while the cyan part for background without semantics is weakened.

inter-channel associations. In general, GDA effectively extracts feature semantics, weakens the non-semantic feature such as background to prevent the interference of complex backgrounds, and provides a robust feature for subsequent target focusing.

3.3 Target Focusing Network (TFN)

The proposed GDA has implemented global semantics modeling in the feature of template and the search region, and separates semantic objects from the non-semantic background. However, it is still unable to distinguish target object and non-target objects both with semantics in the feature. Therefore, in order to achieve accurate target localization, we need to further weaken the feature expression of non-target objects with typical semantics to prevent them from affecting target localization in the feature matching process, which is also the most critical problem. To achieve this, we design the TFN with three steps, as Fig. 6.

First, it builds associations between the template and the search region. Specifically, we take the template feature as query Q and the feature of search region as key K and value V. The point-to-point similarity matrix S between the search region feature $F_s \in \mathbb{R}^{n_s \times d_s}$ and the template feature $F_t \in \mathbb{R}^{n_t \times d_t}$ is defined by following:

$$S = softmax\left(\frac{KQ^T}{\sqrt{d_k}}\right) \tag{5}$$

where d_k is the key dimensionality, $S \in \mathbb{R}^{n_s \times n_t}$ is the similarity matrix between template feature and the search region feature. In this work, we employ $n_t = 25$, $n_s = 625$ and $d_t = d_s = d_k = 256$ as default values.

Second, after obtaining the point-to-point similarity matrix $S \in \mathbb{R}^{n_s \times n_t}$, different from the current popular self attention, we perform a column-by-column (along the dimension representing template) summation within the similarity matrix, and obtain the similarity between each point in the search region feature and the whole target feature. Subsequently, we transform it to the probability that the point belongs to the target through the softmax operation. The process above is given by the following equation:

$$P = softmax(\sum_{j=1}^{n_t} S_{i,j}) \tag{6}$$

where $P \in \mathbb{R}^{n_s \times 1}$ is the probability that each point in the feature of search region belongs to the target. $S_{i,j}$ is the similarity between the point i in the search region feature and point j in the template feature.

Third, after obtaining the P, we perform a pointwise product $(\cdot)$ between P and the output feature N_{input} of search region computed by GDA, and add the result to N_{input} to get the output N_{output} of TFN. The equation for TFN is

$$N_{output} = N_{input} + N_{input} \cdot P \tag{7}$$

The proposed TFN is more resistant to interference than existing feature fusion based approach. Specifically, the methods based on feature fusion take the search region as the query to perform linear process on the template feature, and fuse them to augment the target embedding. However, noise is easily introduced during the information transfer from the search region to the template, which affects the discrimination of the target. In contrast, TFN uses the template as the query and transfers the target information from the template to the search region to build a point-to-point similarity between them, and further achieves part-level target focusing.

We compare the performance between the above two approaches in achieving target focusing. Experiments demonstrate that our TFN has stronger robustness compared to the feature fusion based approach and avoids the influence of background and interferers in the search region. As shown in Fig. 7, compared with the feature fusion based approach, our TFN performs better on highlighting the target from the search region, and has stronger robustness. Such as the third column of Fig. 7, the feature fusion based approach is completely unable to distinguish the tiger toy from background, while our TFN effectively focuses on the area where the tiger toy locates in, and eliminates the distraction of leaves.

Experiment

4.1 Implementation Details

Offline Training. The proposed GdaTFT is implemented in Python on 4 RTX-2080Ti. We use the modified GoogLeNet (Inveption v3) as the backbone for feature extraction. Its pretrained parameters is used as initialization to retrain our model. During the training, the batchsize is set to 96 and totally 20 epochs are performed by using stochastic gradient

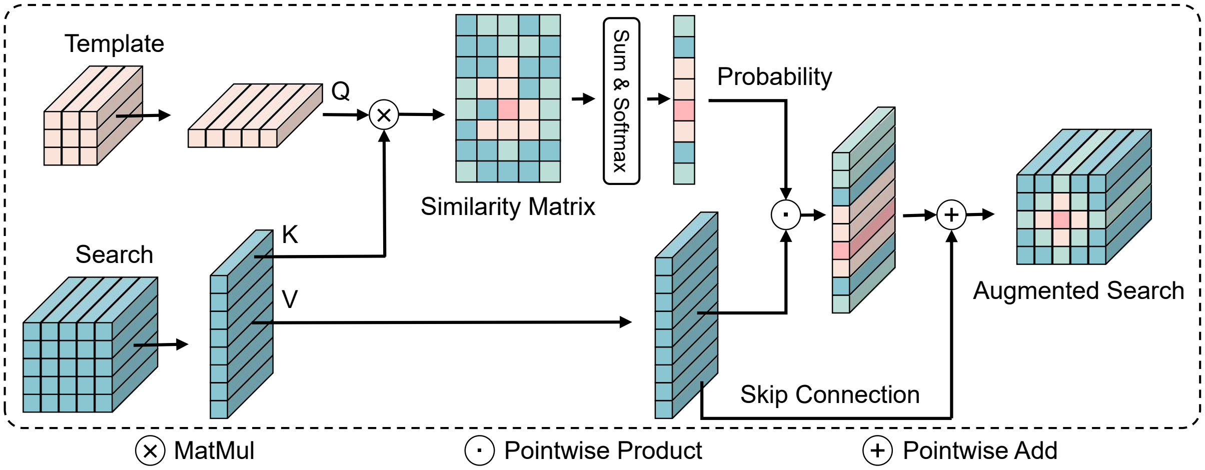


Figure 6: The proposed Target Focus Network (TFN). Using the template as query, the TFN transfers the target information from the template to the search region to augment target embedding, which finally leads to more accurate tracking result.

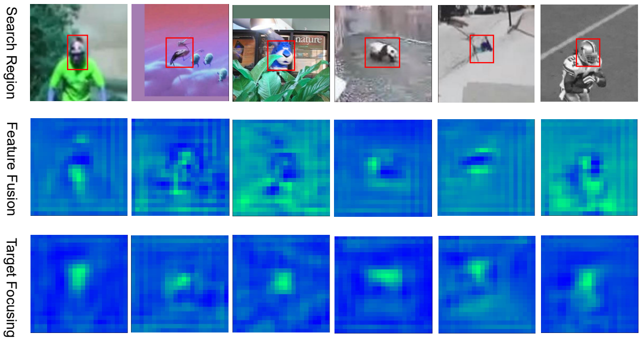


Figure 7: Feature comparisons. Our method produces much more accurate features (third row) than the present feature fusion based methods which often introduce distractors on background and similar objects (second row).

descent. The initial learning rate is 1e-6, which increases linearly to 8e-2 within an epoch, and then decreases to 1e-6 for the rest 19 epochs. We employ three loss of focal loss (Lin et al. 2017), binary cross entropy (De Boer et al. 2005) and IoU loss (Yu et al. 2016) to train the model. We combine the three losses with linear process by the ratio 1:1:2. The whole training phase takes around 72 hours. We train GdaTFN with the data from COCO (Lin et al. 2014), GOT-10k (Huang, Zhao, and Huang 2019), ImageNet DET/VID (Russakovsky et al. 2015), TrackingNet (Muller et al. 2018) and LaSOT (Fan et al. 2019). The patch sizes of search region and template are separately set to 289 × 289 and 127 × 127.

Online Tracking. In online tracking, in order to avoid the influence of distractors, we use the target region from the initial frame as the template, and take it as the input of template. The reason is the initial frame contains the most reliable target without any occlusion and deformation. For the search branch, we expand the predicted target box from the previous frame by a factor to obtain an image of the current search region. In our experiments, we set 4 as the factor.

4.2 Evalution

We compare GdaTFT with state-of-the-art trackers on classic benchmarks, including: GOT-10K(Huang, Zhao, and Huang 2019), OTB100(Wu, Lim, and Yang 2013), TrackingNet(Muller et al. 2018) and LaSOT(Fan et al. 2019). As Table 1, our GdaTFT achieves good performance on various benchmarks. Some visual comparisons are in the appendix.

GOT-10k. GOT-10k(Huang, Zhao, and Huang 2019) is a challenging large-scale real-world tracking benchmark, with a total of 563 types of objects, which requires high generalization capability of the tracker. The test set includes 180 videos. To evaluate the generalization ability of our tracker to real-world scenarios, we compare our GdaTFT with current state-of-the-art trackers on GOT-10k. The results show that we perform very well on this benchmark, where $SR_{0.5}$ is 77.8, surpassing TransT to achieve the current best score, which proves the generalization ability of our GdaTFT.

OTB100. OTB100(Wu, Lim, and Yang 2013) is the most classic benchmark in the tracking field. It covers 11 challenges such as illumination change (IV), deformation (DEF) and occlusion (OCC), and contains 100 test videos in total. We compare the 11 challenges of this benchmark with existing excellent methods, as Fig 8, we are at the highest level in 6 challenges including IV and DEF, surpassing state-of-the-art methods such as TransT, the remaining five are also in the leading position of top-3.

LaSOT. LaSOT(Fan et al. 2019) is a large long-term tracking dataset with 1400 videos covering various tracking challenges, with 280 videos in the test set. Each video has an average of 2512 frames, and the longest contains 11397 frames, which is a great test of the tracker's robustness and long-term tracking capability. The results show that on this benchmark, our method outperforms most existing trackers and is on par with state-of-the-art methods such as TransT (Chen et al. 2021), which demonstrates that our method is highly competitive in long-term tracking scenarios.

TrackingNet. TrackingNet(Muller et al. 2018) is a large-scale short-term tracking benchmark that provides 511

Tracker	LaSOT			TrackingNet			GOT-10k		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.75}$	$SR_{0.5}$
TransT (Chen et al. 2021)	**64.9**	**73.8**	**69.0**	**81.4**	**86.7**	**80.3**	**67.1**	**60.9**	**76.8**
TrDimp (Wang et al. 2021)	**63.9**	-	61.4	**78.4**	**83.3**	**73.1**	**67.1**	**58.3**	**77.7**
SiamGAT (Guo et al. 2021)	53.9	53.0	**63.3**	-	-	-	62.7	48.8	74.3
CGACD (Du et al. 2020)	51.8	62.6	-	71.1	80.0	69.3	-	-	-
Ocean-online (Zhang et al. 2020)	56.0	**65.1**	56.6	-	-	-	61.1	47.3	72.1
Ocean-offline (Zhang et al. 2020)	52.6	-	52.6	-	-	-	59.2	-	69.5
SiamFC++ (Xu et al. 2020)	54.3	54.7	62.3	75.4	80.0	70.5	59.5	47.9	69.5
SiamCAR (Guo et al. 2020)	50.7	60.0	51.0	-	-	-	56.9	41.5	67.0
FCOT (Cui et al. 2020)	57.2	67.8	-	75.4	82.9	72.6	63.4	52.1	76.6
SiamAttn (Yu et al. 2020)	56.0	64.8	-	75.2	81.7	-	-	-	-
DiMP50 (Bhat et al. 2019)	56.9	65.0	56.7	74.0	80.1	68.7	61.1	49.2	71.7
ATOM (Danelljan et al. 2019)	51.5	57.6	50.5	70.3	77.1	64.8	55.6	63.4	40.2
SiamRPN++ (Li et al. 2019)	49.6	56.9	49.1	73.3	80.0	69.4	51.7	32.5	61.6
GdaTFT (Ours)	**64.3**	**68.0**	**68.7**	**77.8**	**83.5**	**75.4**	**65.0**	**53.7**	**77.8**

Table 1: Comparisions on TrackingNet, LaSOT, GOT-10k. The top-3 results are shown in **red**, **blue** and green fonts.

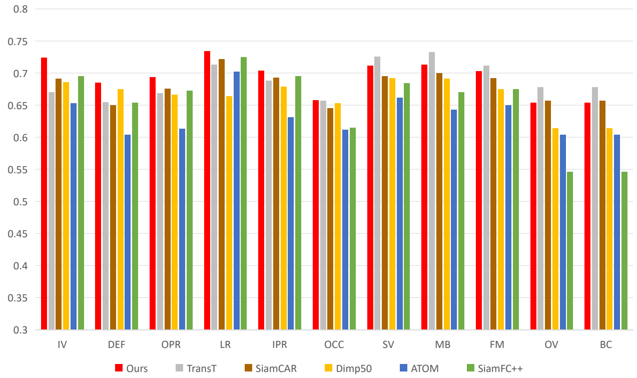


Figure 8: Comparisons on OTB100 with 11 challenges, including Illumination Variation (IV), Deformation (DEF), Out-of-Plane Rotation (OPR), Low Resolution (LR), In-Plane Rotation (IPR), Occlusion (OCC), Scale Variation (SC), Motion Blur (MB), Fast Motion (FM), Out of View (OV), and Background Clutter (BC).

videos without published groundtruth for testing. We upload the tracking results of our method to the official evaluation server to obtain its performance on the three indicators of AUC, Precision (P) and Normalized Precision (P_{Norm}). Results show that our method outperforms most existing tracking methods and is second only to TransT (Chen et al. 2021).

4.3 Ablation Study

Comparison with the current popular self attention. To investigate the effectiveness of our GDA, we trained two models using GDA and self attention (SA) respectively with the same framework and tested them on OTB100. As in Table 2, the results show that both self attention and GDA can improve the tracking performance, but our GDA is 1.2% and 0.7% higher than self attention on Precision and Success, which proves the effectiveness of our GDA.

Comparison with the feature fusion based approach. To conduct a comparison between feature fusion (FF) based approach and our TFN, we trained two models using TFN and feature fusion based modules respectively and tested them on OTB100. As in Table 2, The results show that our TFN is 1.1% and 0.6% higher than feature fusion based approach on Precision and Success, which proves the effectiveness of our TFN. Furthermore, The results show that the best performance can be obtained when GDA and TFN are used together, which also justifies the effectiveness of our GdaTFT.

SA	GDA	FF	TFN	OTB100	
				Success	Precision
✗	✗	✗	✗	67.5	87.4
✓	✗	✗	✗	68.3 (↑ 0.8)	88.3 (↑ 0.9)
✗	✓	✗	✗	69.0 (↑ 1.5)	89.5 (↑ 2.1)
✗	✗	✓	✗	68.6 (↑ 1.1)	88.4 (↑ 1.0)
✗	✗	✗	✓	69.2 (↑ 1.7)	89.5 (↑ 2.1)
✗	✓	✗	✓	**70.4 (↑ 2.9)**	**90.6 (↑ 3.2)**

Table 2: Ablation study: experiment resutls on OTB100. The best performance (last row) is obtained with both the proposed GDA and TFN.

5 Conclusion

In this paper, we propose a novel tracking framework named GdaTFT for general object tracking by defining the GDA and TFN. The proposed GDA designs a dilated attention based on multi-head self attention and dilated split and concatenation, and introduces local focusing module and channel adaption module to enhance the semantics of feature. the TFN takes the template as the query to build a point-point similarity between the tempalte and search region, and finally achieves the part-level enhancement of target embedding. Experiments on various challenging benchmarks including GOT-10K, OTB100, LaSOT and TrackingNet show that our GdaTFT outperforms many start-of-the-art trackers and achieves leading performance.

Acknowledgments

This project was supported by the National Natural Science Foundation of China (61772209), the Key Research and Development project of Guangzhou (202206010091), the Science and Technology Planning Project of Guangdong Province (2019A050510034).

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, 850–865. Springer.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6182–6191.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8126–8135.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6668–6677.

Choi, J.; Jin Chang, H.; Yun, S.; Fischer, T.; Demiris, Y.; and Young Choi, J. 2017. Attentional correlation filter network for adaptive visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4807–4816.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2020. Fully convolutional online tracking. *arXiv preprint arXiv:2004.07109*.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2021. Target transformed regression for accurate tracking. *arXiv preprint arXiv:2104.00403*.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4660–4669.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1): 19–67.

Du, F.; Liu, P.; Zhao, W.; and Tang, X. 2020. Correlation-guided attention for corner detection based visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6836–6845.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5374–5383.

Fu, Z.; Liu, Q.; Fu, Z.; and Wang, Y. 2021. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13774–13783.

Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; and Shen, C. 2021. Graph attention tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9543–9552.

Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; and Chen, S. 2020. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6269–6277.

Huang, L.; Zhao, X.; and Huang, K. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1562–1577.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4282–4291.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8971–8980.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, 300–317.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1571–1580.

Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; and Maybank, S. 2018. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4854–4863.

Wu, Y.; Lim, J.; and Yang, M.-H. 2013. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2411–2418.

Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; and Yu, G. 2020. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12549–12556.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10448–10457.

Yu, B.; Tang, M.; Zheng, L.; Zhu, G.; Wang, J.; Feng, H.; Feng, X.; and Lu, H. 2021. High-performance discriminative tracking with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9856–9865.

Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; and Huang, T. 2016. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, 516–520.

Yu, Y.; Xiong, Y.; Huang, W.; and Scott, M. R. 2020. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6728–6737.

Zhang, Z.; Peng, H.; Fu, J.; Li, B.; and Hu, W. 2020. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, 771–787. Springer.