

Reidentifikácia vozidiel v snímkach z dopravných kamier

Autor: Richard Dominik
FMFI UK 16.12.2021

www.st.fmph.uniba.sk/~dominik3/



The Devil is in the Details: Self-Supervised Attention for Vehicle Re-Identification

Pirazh Khorramshahi*¹, Neehar Peri*¹, Jun-cheng Chen², and Rama Chellappa¹

¹ Center for Automation Research, UMIACS, and the Department of Electrical and Computer Engineering, University of Maryland, College Park

² Research Center for Information Technology Innovation, Academia Sinica

Abstract. In recent years, the research community has approached the problem of vehicle re-identification (re-id) with attention-based models, specifically focusing on regions of a vehicle containing discriminative information. These re-id methods rely on expensive key-point labels, part annotations, and additional attributes including vehicle make, model, and color. Given the large number of vehicle re-id datasets with various levels of annotations, strongly-supervised methods are unable to scale across different domains. In this paper, we present Self-supervised Attention for Vehicle Re-identification (SAVER), a novel approach to effectively learn vehicle-specific discriminative features. Through extensive experimentation, we show that SAVER improves upon the state-of-the-art on challenging VeRi, VehicleID, Vehicle-1M and VERI-Wild datasets.

Keywords: Vehicle Re-Identification, Self-Supervised Learning, Variational Auto-Encoder, Deep Representation Learning

1 Introduction



This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.

Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Bag of Tricks and A Strong Baseline for Deep Person Re-identification

Hao Luo^{1*}, Youzhi Gu^{1*}, Xingyu Liao^{2*}, Shenqi Lai³, Wei Jiang¹

¹Zhejiang University, ²Chinese Academy of Sciences, ³Xi'an Jiaotong University

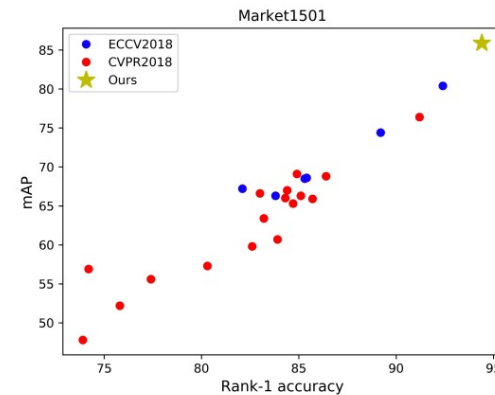
{haoluocsc, gu_youzhi, jiangwei_zju}@zju.edu.cn randall@mail.ustc.edu.cn laishenqi@stu.xjtu.edu.cn

Abstract

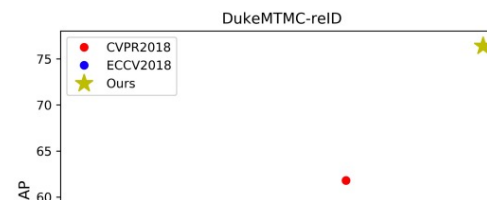
This paper explores a simple and efficient baseline for person re-identification (ReID). Person re-identification (ReID) with deep neural networks has made progress and achieved high performance in recent years. However, many state-of-the-arts methods design complex network structure and concatenate multi-branch features. In the literature, some effective training tricks are briefly appeared in several papers or source codes. This paper will collect and evaluate these effective training tricks in person ReID. By combining these tricks together, the model achieves 94.5% rank-1 and 85.9% mAP on Market1501 with only using global features. Our codes and models are available at <https://github.com/michuanhaohao/reid-strong-baseline>

1. Introduction

Person re-identification (ReID) with deep neural networks has made progress and achieved high performance in recent years. However, many state-of-the-arts methods



(a) Market1501



O čom bude dnešná prezentácia ?

Problém a využitie

{1}

{2}

Datasets

Riešenie autorov

{3}

Metriky a výsledky

{4}

Nápady na vylepšenie

{6}

{7}

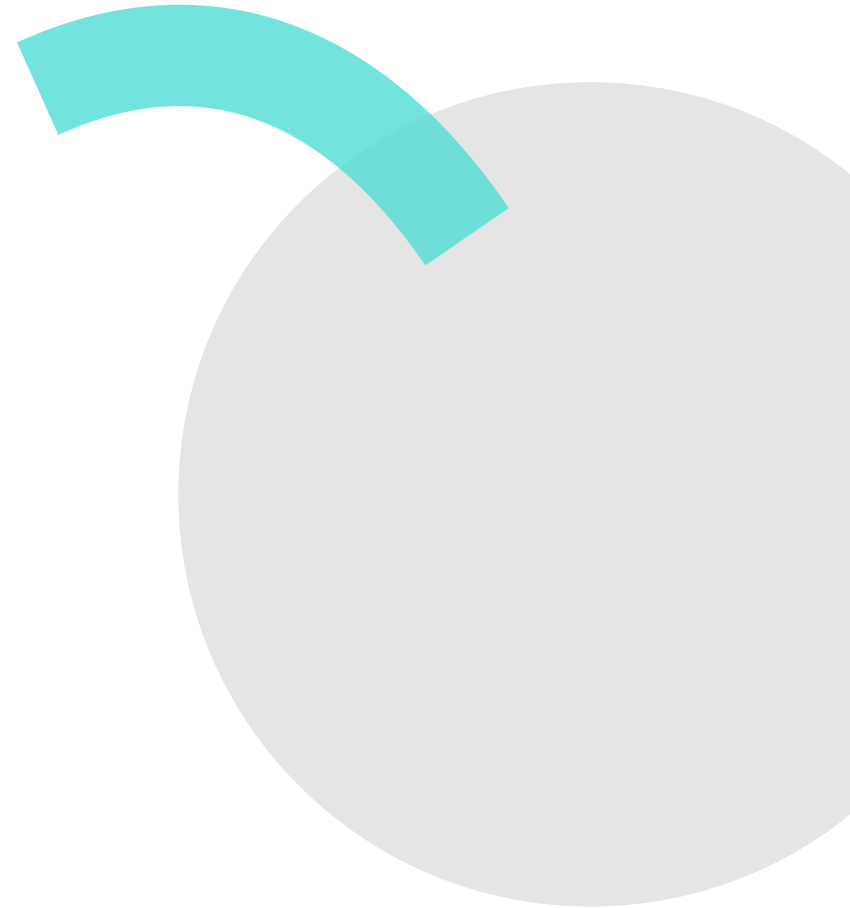
State of the art

Problém a využitie

- zhoda rovnakého vozidla na snímkach z veľkého datasetu obrázkov (nasnímané pomocou dopravných kamier)
- rôzne kamery, orientácie, čas, lokácie, oklúzie, nezaostrenosť ...
- podobný tvar, model, farba, výrobca...
- reidentifikácia vozidiel != detekcia vozidiel

Problém a využitie

- aktuálna téma v oblasti počítačového videnia
- príbuzné k téme reidentifikácie osôb
- vieme identifikovať auto v rôznych bodoch mesta (odkiaľ kam išlo)
- využitie v inteligentných dopravných systémoch (efektívnejšie navrhovanie dopravných sietí)



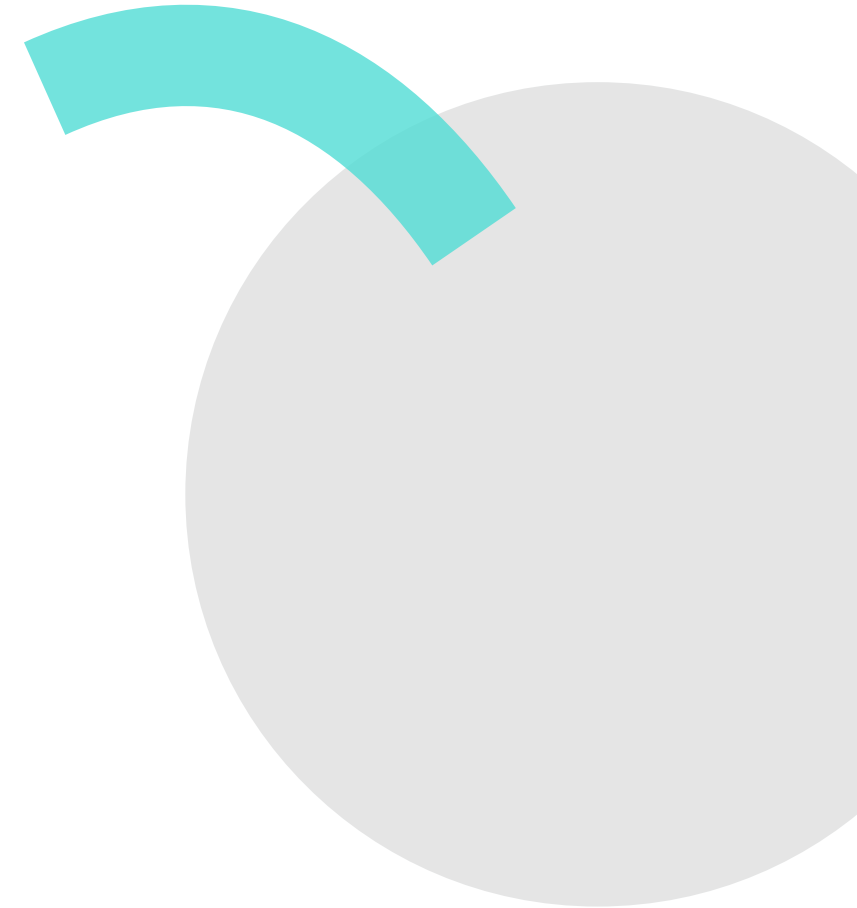
Náročnejšie prípady ?!



Obrázok: Ukážka prípadov zlyhania reidentifikácie (a, b rovnaký smer) a (c, d podobné pozadie)

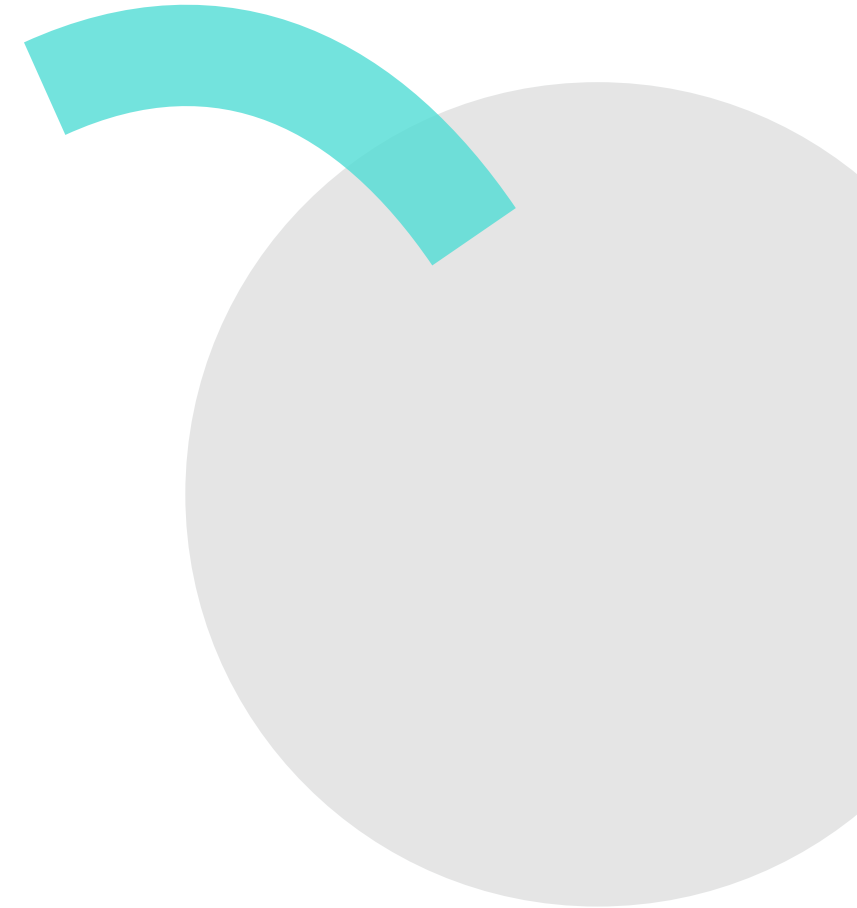
Datase~~ty~~

- **AI City Challenge dataset**
- **VeRi-776**
- Stanford Cars
- VERI – Wild
- CompCars
- VRAI
- Vehicle-1M
- VehicleX
- VehicleID
- BoxCars 116K



AI City Challenge dataset

- dáta nasnímané z dopravných kamier v USA (štát Iowa)
- 85 058 obrázkov
- 52 717 trénovacích a 31 238 testovacích obrázkov
- 440 rôznych vozidiel
- nasnímané pomocou 46 kamier
- anotované ľuďmi (farba, model, typ...)

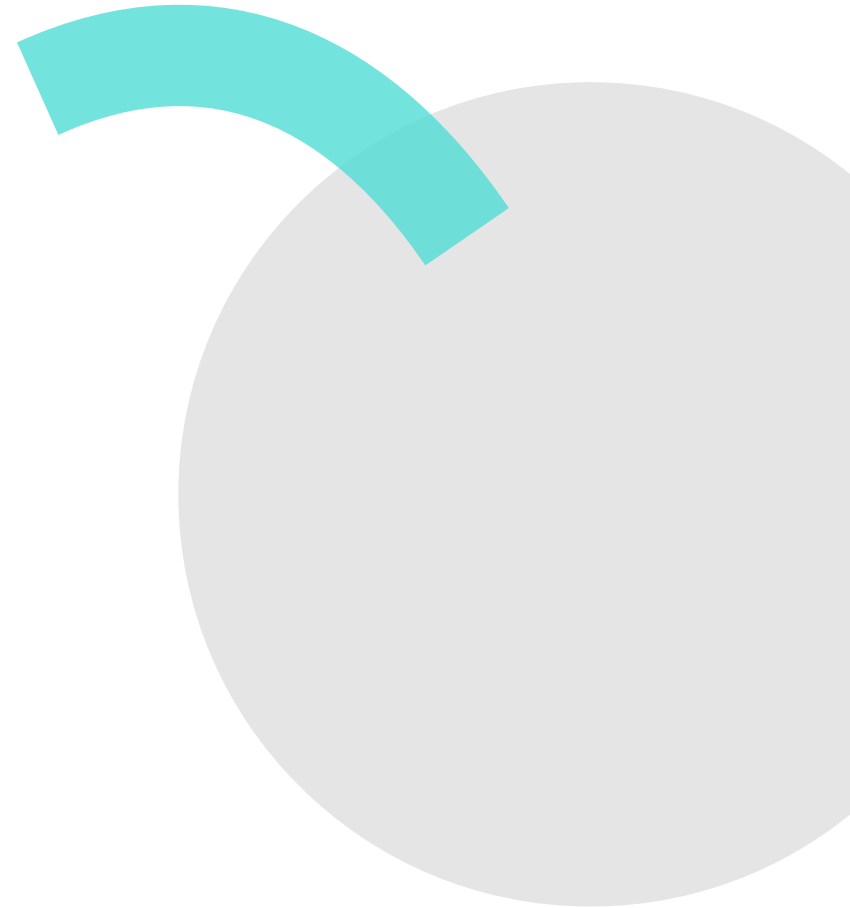


AI City Challenge dataset



VeRi-776 dataset

- 50 000+ obrázkov
- 776 rôznych vozidiel
- nasnímané pomocou 20 kamier
- rôzne pohľady, rozlíšenia, svetelné podmienky, oklúzie
- anotácie (Bbox, typ, farba, značka)

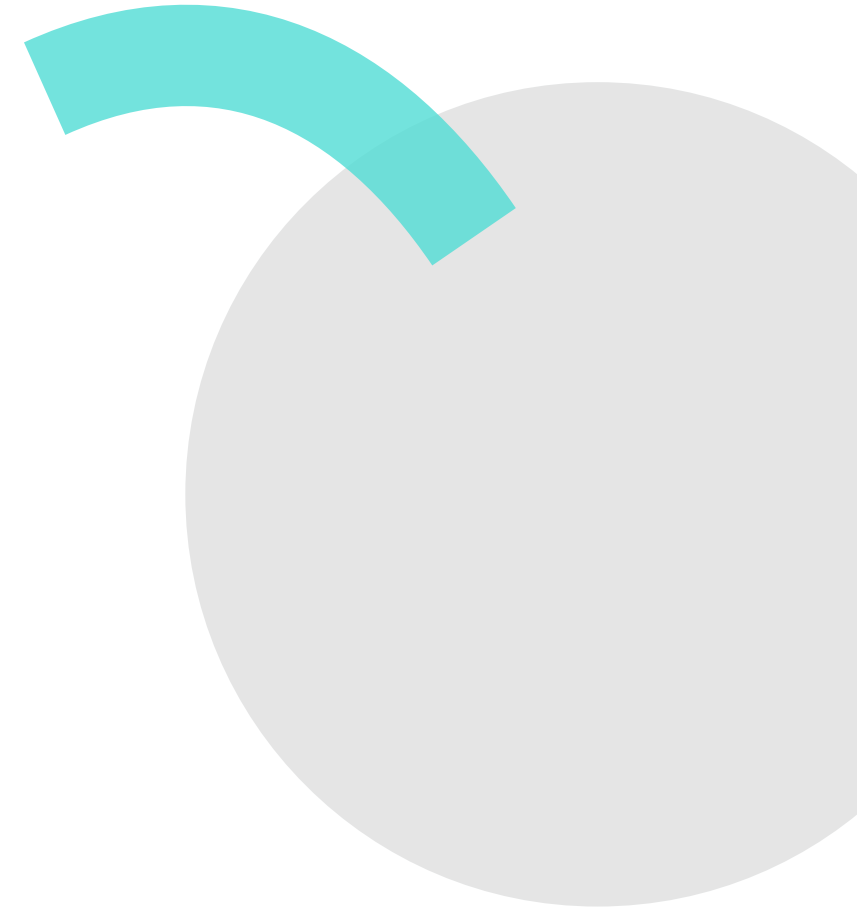


VeRI-776 dataset



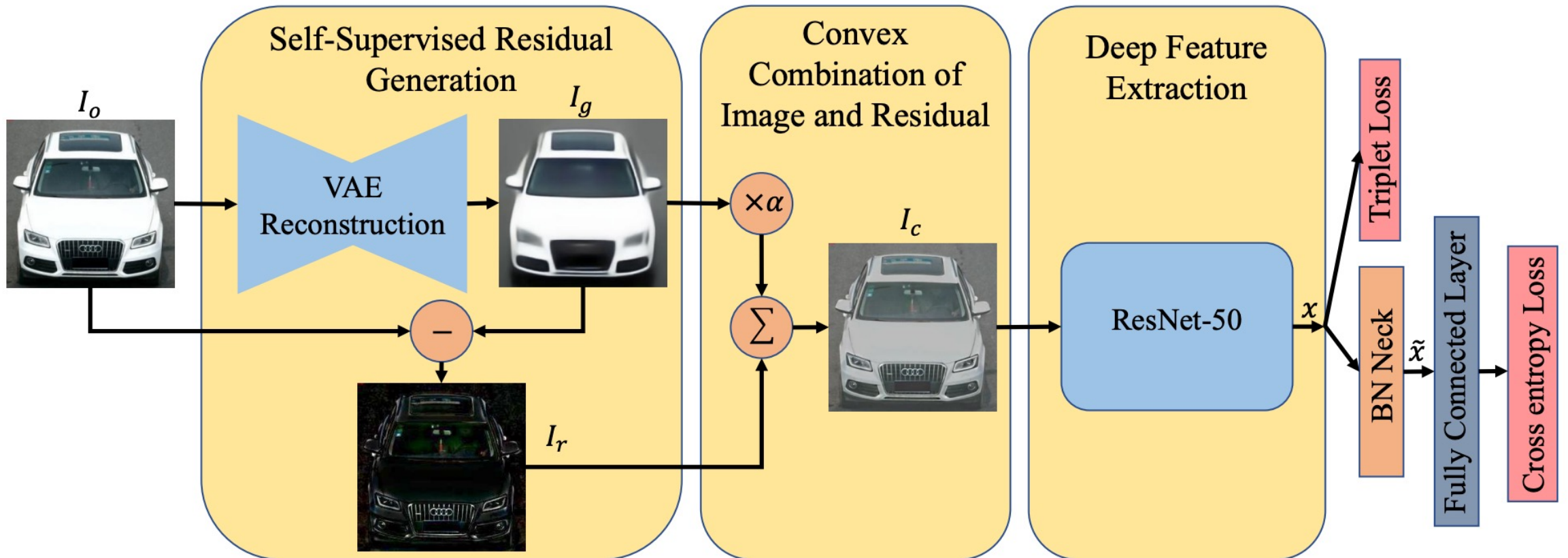
Prístup autorov

- zmenená veľkosť obrázkov na rozmer 256x256 a následná normalizácia
- Adam optimalizátor
- 150 epôch
- VAE (Variational Auto-Encoder)
- ResNet 50 pre extrakciu príznakov
- Triplet Loss a Cross entropy Loss
- Trénovacie triky (Warmup learning rate, REA...)



Architektúra

$$I_c = \alpha \times I_o + (1 - \alpha) \times I_r$$



Trénovacie triky*

*inšpirované publikáciou venovanej reidentifikácii osôb

Trénovacie triky

- Warmup learning rate (t je číslo epochy)

$$lr(t) = \begin{cases} 3.5 \times 10^{-5} \times \frac{t}{10} & \text{if } t \leq 10 \\ 3.5 \times 10^{-4} & \text{if } 10 < t \leq 40 \\ 3.5 \times 10^{-5} & \text{if } 40 < t \leq 70 \\ 3.5 \times 10^{-6} & \text{if } 70 < t \leq 120 \end{cases}$$

- Random Erasing Augmentation (REA)



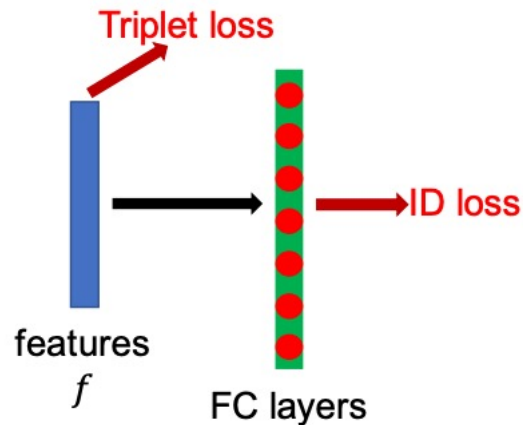
Trénovacie triky

- Label smoothing

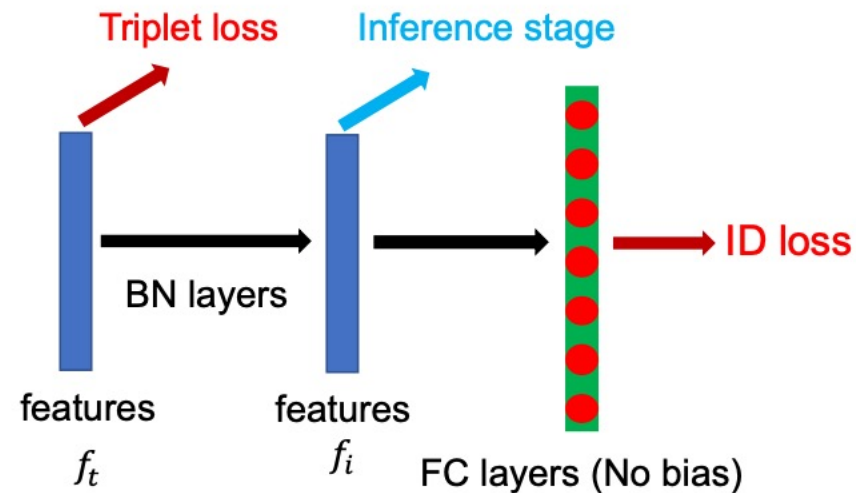
$$L(ID) = \sum_{i=1}^N -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases}$$

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\epsilon & \text{if } i = y \\ \epsilon/N & \text{otherwise,} \end{cases} \quad \epsilon = 0.1$$

- Batch normalization (BN) Neck



(a) The neck of the standard baseline.



(b) Our designed BNNeck. In the inference stage, we choose f_i following the BN layer to do the retrieval.

Aké príznaky sú pri reidentifikácii vozidla podľa predstaveného prístupu dôležité ?



Metriky a výsledky na datasete VeRi 776

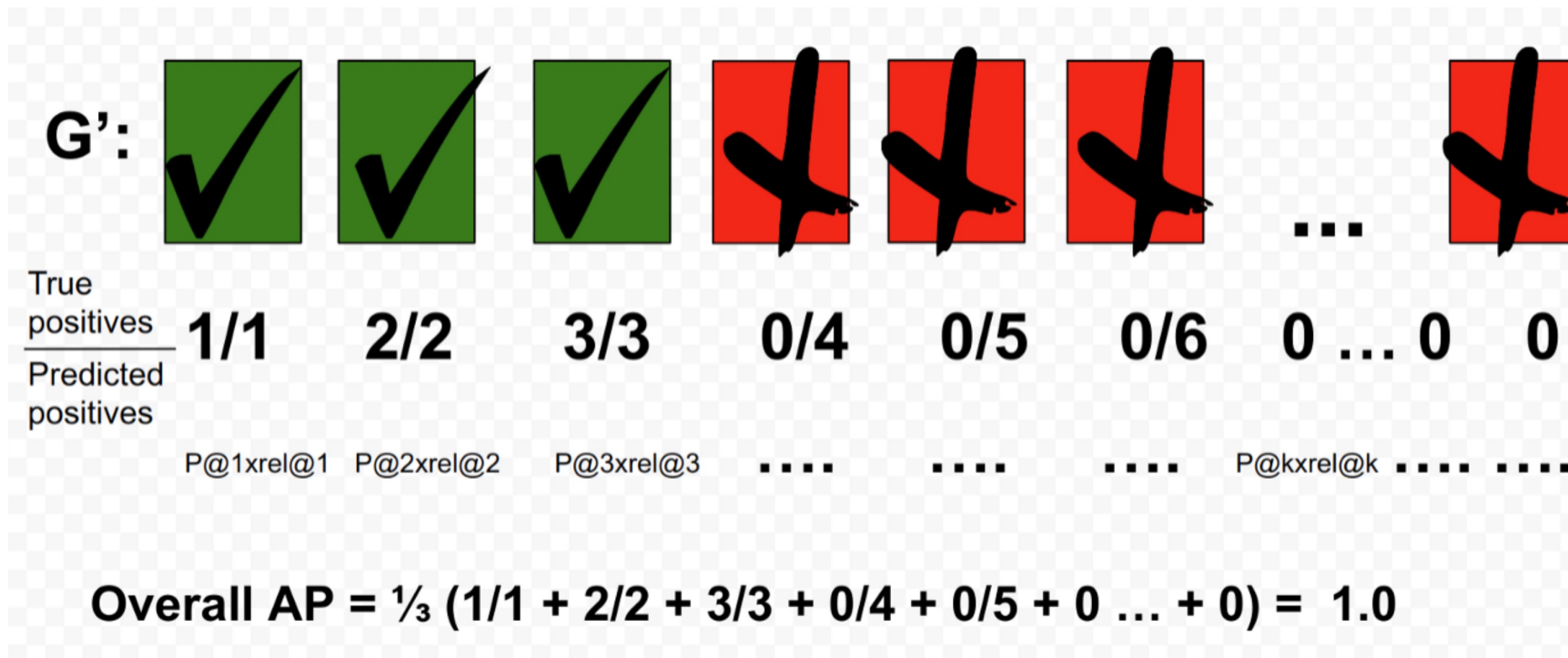
- Mean Average Precision

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{AP@}n = \frac{1}{\text{GTP}} \sum_k^n \text{P@}k \times \text{rel@}k$$

Metriky a výsledky na datasete VeRi 776



$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i$$

Metriky a výsledky na datasete VeRi 776

- v publikácii dosiahli výsledok **79.6% mAP**
- tento výsledok sa autorom ešte pomocou metódy re-ranking podarilo vylepšiť na **82% mAP**

Nápady na vylepšenie ?

Tranformery

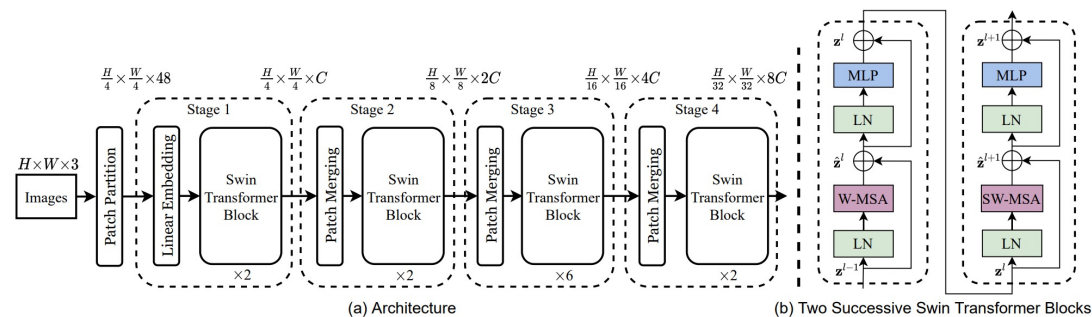
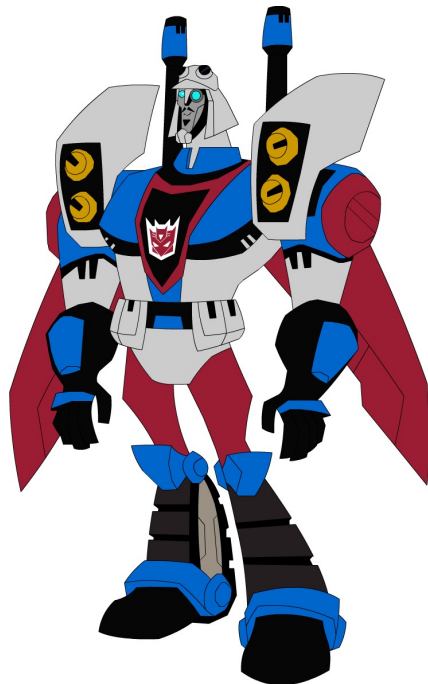


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.



ViT | Swin-Transformer (v2 ?)

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy^{*†}, Lucas Beyer^{*}, Alexander Kolesnikov^{*}, Dirk Weissenborn^{*},
Xiaohua Zhai^{*}, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby^{*†}
^{*}equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.¹

1 INTRODUCTION

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu^{†*} Yutong Lin^{†**} Yue Cao^{*} Han Hu^{**‡} Yixuan Wei[†]
Zheng Zhang Stephen Lin Baining Guo
Microsoft Research Asia

{v-zeliul, v-yutlin, yuecao, hanhu, v-yixwe, zhez, stevelin, bainguo}@microsoft.com

Abstract

This paper presents a new vision Transformer, called Swin Transformer, that capably serves as a general-purpose backbone for computer vision. Challenges in adapting Transformer from language to vision arise from differences between the two domains, such as large variations in the scale of visual entities and the high resolution of pixels in images compared to words in text. To address these differences, we propose a hierarchical Transformer whose representation is computed with Shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to image size. These qualities of Swin Transformer make it compatible with a broad range of vision tasks, including image classification (87.3 top-1 accuracy on ImageNet-1K) and dense prediction tasks such as object detection (58.7 box AP and 51.1 mask AP on COCO test-dev) and semantic segmentation (53.5 mIoU on ADE20K val). Its performance surpasses the previous state-of-the-art by a large margin of +2.7 box AP and +2.6 mask AP on COCO, and +3.2 mIoU on ADE20K, demonstrating the potential of Transformer-based models as vision backbones. The hierarchical design and the shifted window approach also prove beneficial for all MLP architectures. The code

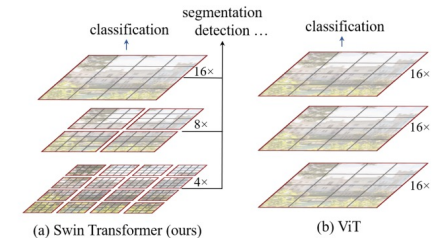


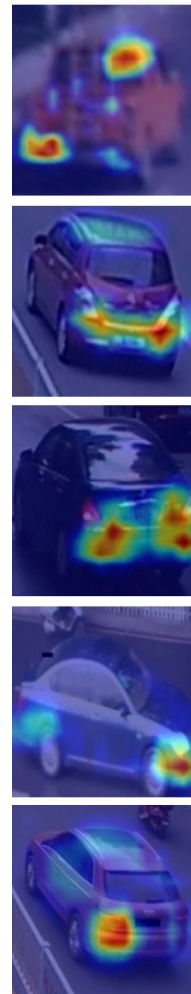
Figure 1. (a) The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous vision Transformers [20] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

greater scale [30, 76], more extensive connections [34], and more sophisticated forms of convolution [70, 18, 84]. With CNNs serving as backbone networks for a variety of vision tasks, these architectural advances have led to performance improvements that have broadly lifted the entire field.

CNN vs Transformer



(a)



(b)















(d)

Obrázok: Grad-cam a) pôvodný obrázok, b) CNN backbone, d) Transformer (ViT) backbone

State of the art pri detekcii objektov COCO test-dev

Rank	Model	box↑ AP	AP50	AP75	APS	APM	APL	Extra Training Data	Paper	Code	Result	Year	Tags
1	SwinV2-G (HTC++)	63.1						✓	Swin Transformer V2: Scaling Up Capacity and Resolution			2021	swin-transformer
2	Florence-CoSwin-H	62.4						✓	Florence: A New Foundation Model for Computer Vision			2021	
3	Soft Teacher + Swin-L (HTC++, multi-scale)	61.3						✓	End-to-End Semi-Supervised Object Detection with Soft Teacher			2021	multiscale Swin-Transformer
4	DyHead (Swin-L, multi scale, self-training)	60.6	78.5	66.6		64.0	74.2	×	Dynamic Head: Unifying Object Detection Heads with Attentions			2021	multiscale Swin-Transformer
5	Dual-Swin-L (HTC, multi-scale)	60.1						×	CBNetV2: A Composite Backbone Network Architecture for Object			2021	multiscale Swin-Transformer

State of the reidentifikácia vozidiel Veri 776

Rank	Model	mAP↑	Rank-1	Rank1	Rank5	Extra Training Data	Paper	Code	Result	Year	Tags 
1	RPTM	87.4	96.2	96.2	98.1	×	Relation Preserving Triplet Mining for Stabilizing the Triplet Loss in Vehicle Re-identification			2021	
2	A Strong Baseline	87.1				×	A Strong Baseline for Vehicle Re-Identification			2021	
3	vehiclenet	83.41	96.78			✓	VehicleNet: Learning Robust Feature Representation for Vehicle Re-identification			2020	
4	TransReID	82.3	97.1			×	TransReID: Transformer-based Object Re-Identification			2021	
5	ANet	81.2	96.8	96.8	98.4	×	AttributeNet: Attribute Enhanced Vehicle Re-Identification			2021	
6	CAL	74.3	95.4		97.9	×	Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification			2021	
7	QD-DLF	61.83				×	Vehicle Re-identification Using Quadruple Directional Deep Learning Features			2018	

Zdroje

- Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa **The Devil is in the Details: Self-Supervised Attention for Vehicle Re-Identification**
(<https://arxiv.org/pdf/2004.06271.pdf>)
- Hao Luo , Youzhi Gu , Xingyu Liao , Shenqi Lai, Wei Jiang **Bag of Tricks and A Strong Baseline for Deep Person Re-identification**
(https://openaccess.thecvf.com/content_CVPRW_2019/papers/TRMTMCT/Luo_Bag_of_Tricks_and_a_Strong_Baseline_for_Deep_Person_CVPRW_2019_paper.pdf)
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo **Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**
(<https://arxiv.org/pdf/2102.04378.pdf>)
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale** (<https://arxiv.org/pdf/2010.11929.pdf>)

Zdroje

- Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, Wei Jiang **TransReID: Transformer-based Object Re-Identification** (<https://arxiv.org/pdf/2103.14030.pdf>)
- <https://www.aicitychallenge.org/2021-data-and-evaluation/>
- <https://github.com/JDAI-CV/VeRidataset>
- <https://paperswithcode.com/sota/object-detection-on-coco>
- <https://paperswithcode.com/sota/vehicle-re-identification-on-veri-776>
- <https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52>



Ďakujem za pozornosť !



Q&A