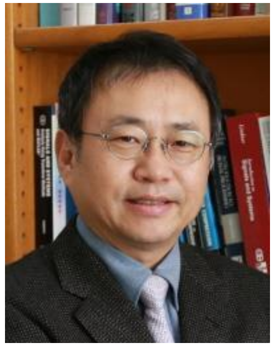
Neural Information
Processing Systems
(NeurIPS) 2018

Recurrent Transformer Networks for Semantic Correspondence







Seungryong Kim[1], Stepthen Lin[2], Sangryul Jeon[1], Dongbo Min[3], Kwanghoon Sohn[1]

Dec. 05, 2018

1) YONSEI UNIVERSITY 2) Microsoft Research 3) 이화여자대학교 EWHA WOMANS UNIVERSITY

Introduction

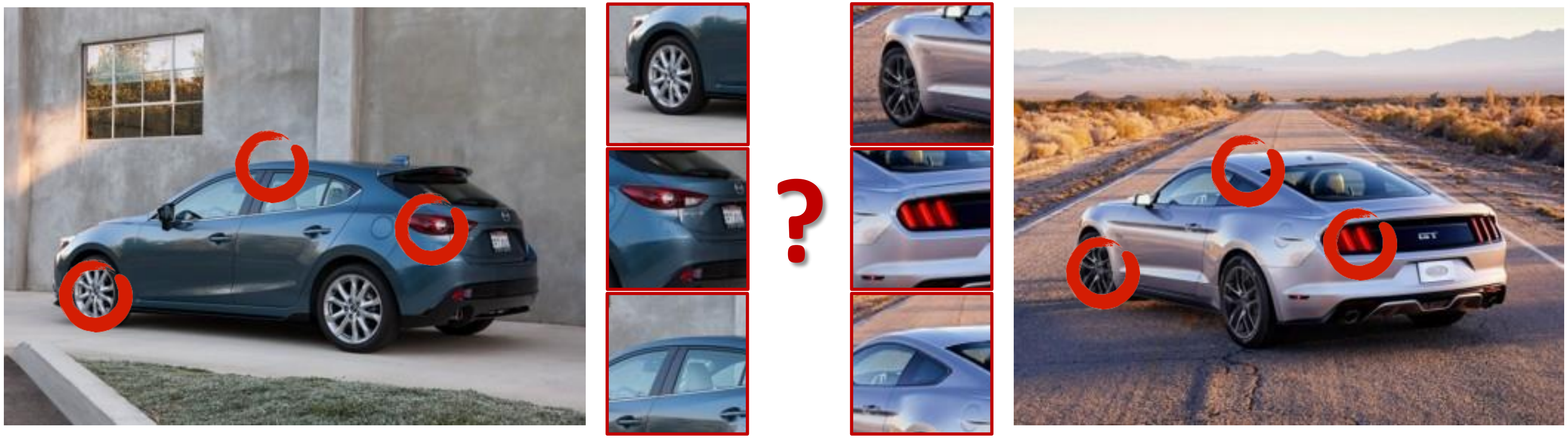
Semantic Correspondence

- Establishing **dense correspondences** between **semantically similar images**, i.e., different instances within the same object or scene categories
- *For example, the wheels of two different cars, the body of people or animals*

Introduction

Challenges in Semantic Correspondence



Photometric Deformations

- *Intra-class appearance and attribute variations*
- *Etc.*

Geometric Deformations

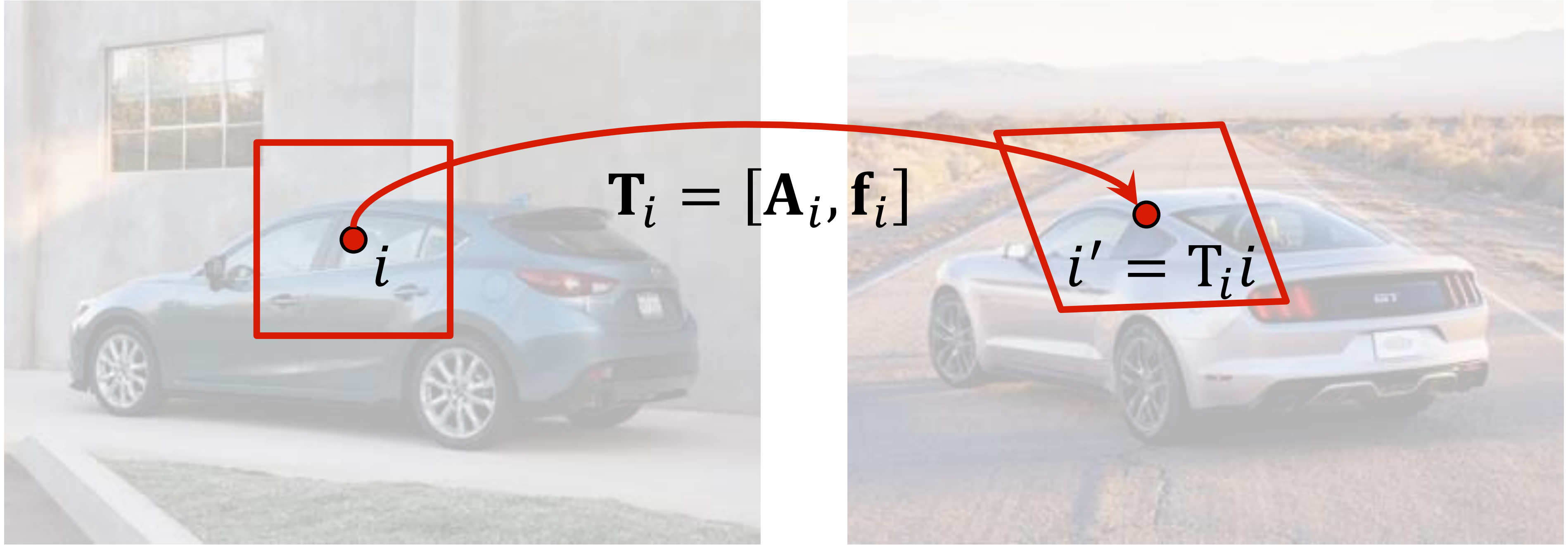
- *Different viewpoint or baseline*
- *Non-rigid shape deformations*
- *Etc.*

Lack of Supervision

- *Labor-intensive of annotation*
- *Degraded by subjectivity*
- *Etc.*

Problem Formulation

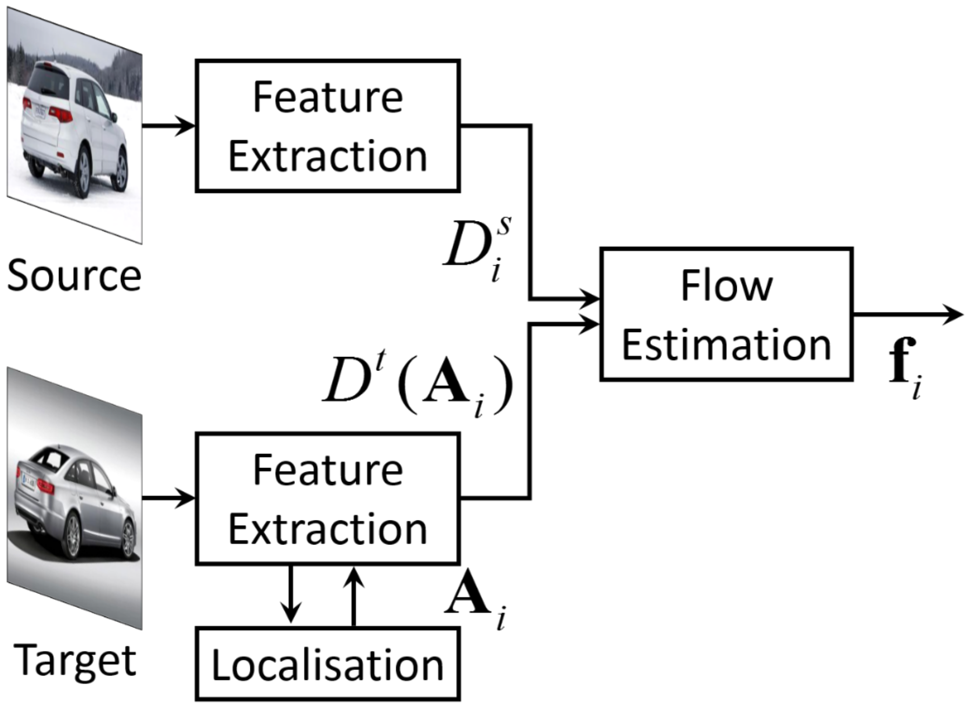
Objective



How to estimate <u>locally-varying affine transformation fields</u> without ground-truth supervision?

Background

Methods for Geometric Invariance in Feature Extraction Step



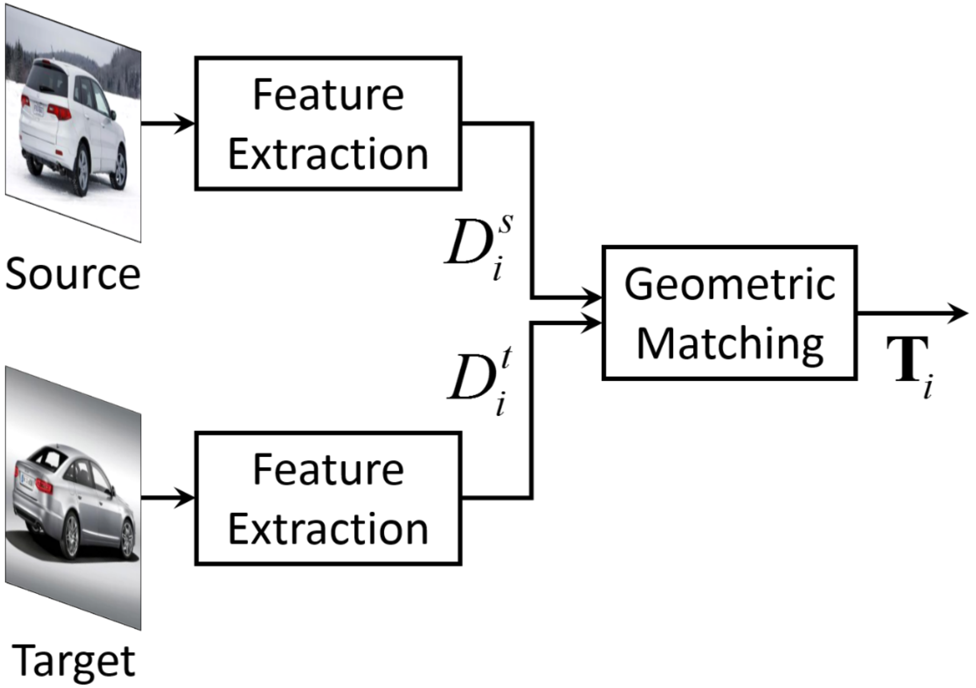
- **UCN** [Choy *et al.*, NeurIPS'16]
- **CAT-FCSS** [Kim *et al.*, TPAMI'18]
- Etc.

✓ Spatial Transformer Networks (STNs)-based methods [Jaderberg et al., NeurIPS'15]
✓ $\mathbf{A}_i$ is learned wo/$\mathbf{A}_i^*$

✗ But, $\mathbf{f}_i$ is learned w/$\mathbf{f}_i^*$
✗ Geometric inference based on only source or target image

Background

Methods for Geometric Invariance in Regularization Step



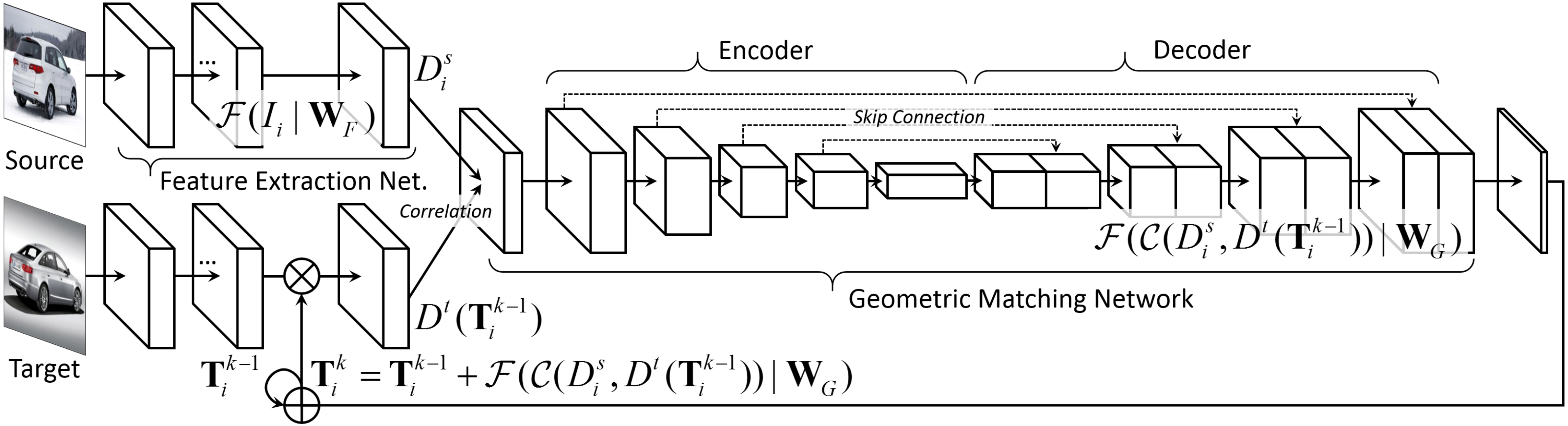
- ✓ $\mathbf{T}_i$ is learned wo/$\mathbf{T}_i^*$ using self- or meta-supervision
- ✓ Geometric Inference using source/target images

- ✗ Globally-varying geometric Inference only
- ✗ Only fixed, untransformed versions of the features

- **GMat.** [Rocco *et al.*, CVPR'17]
- **GMat. w/Inl.** [Rocco *et al.*, CVPR'18]
- Etc.

Recurrent Transformer Networks (RTNs)

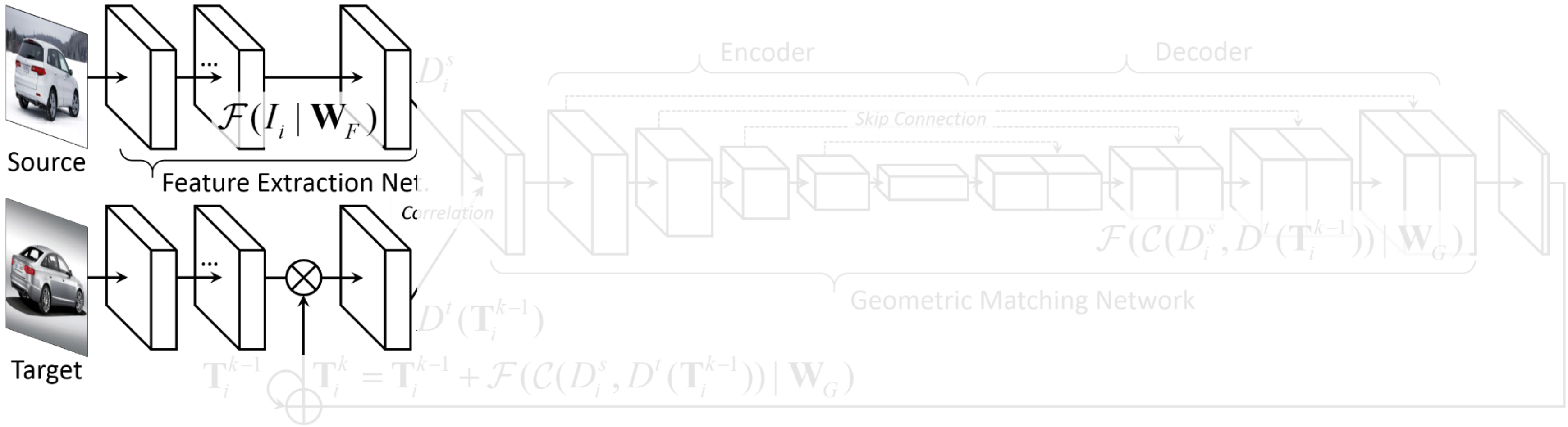
Networks Configuration



- To weaves the advantages of **STN-based methods** and **geometric matching methods** by recursively estimating geometric transformation residuals using geometry-aligned feature activations

Recurrent Transformer Networks (RTNs)

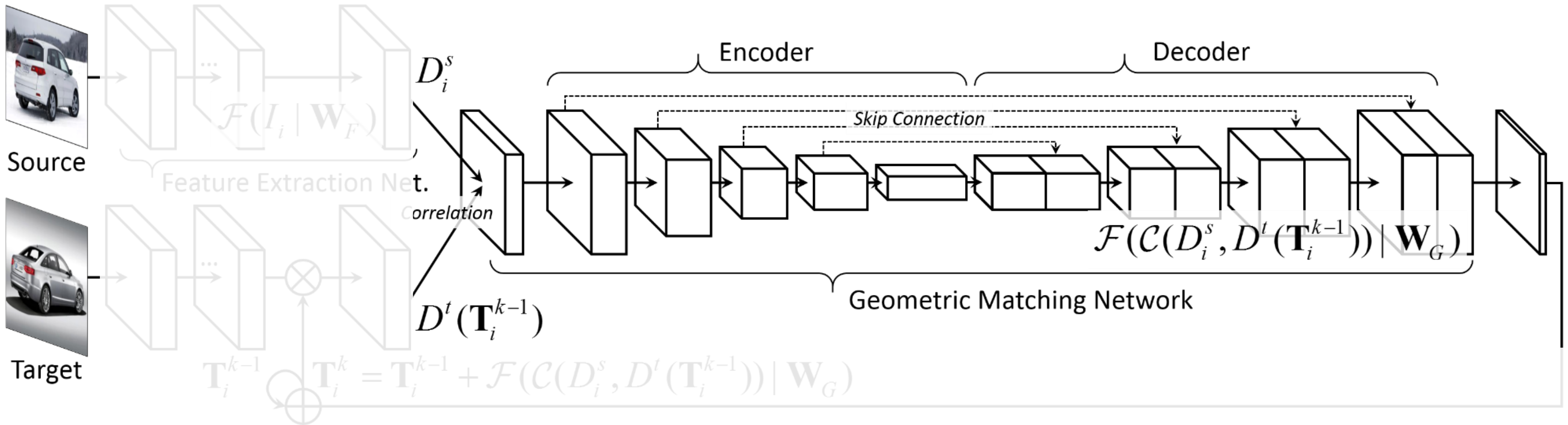
Feature Extraction Networks



- Input images I^s and I^t are passed through Siamese convolution networks with parameters $\mathbf{W}_F$ such that $D_i = F(I|\mathbf{W}_F)$
- Using CAT-FCSS, VGGNet (conv4-4), ResNet (conv4-23)

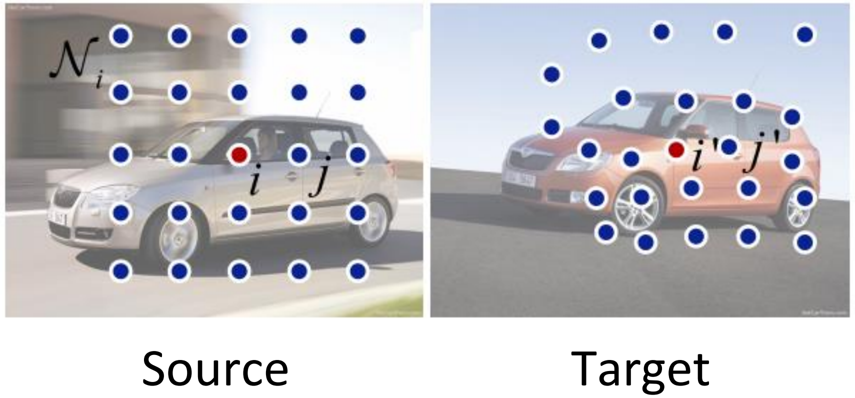
Recurrent Transformer Networks (RTNs)

Recurrent Geometric Matching Networks



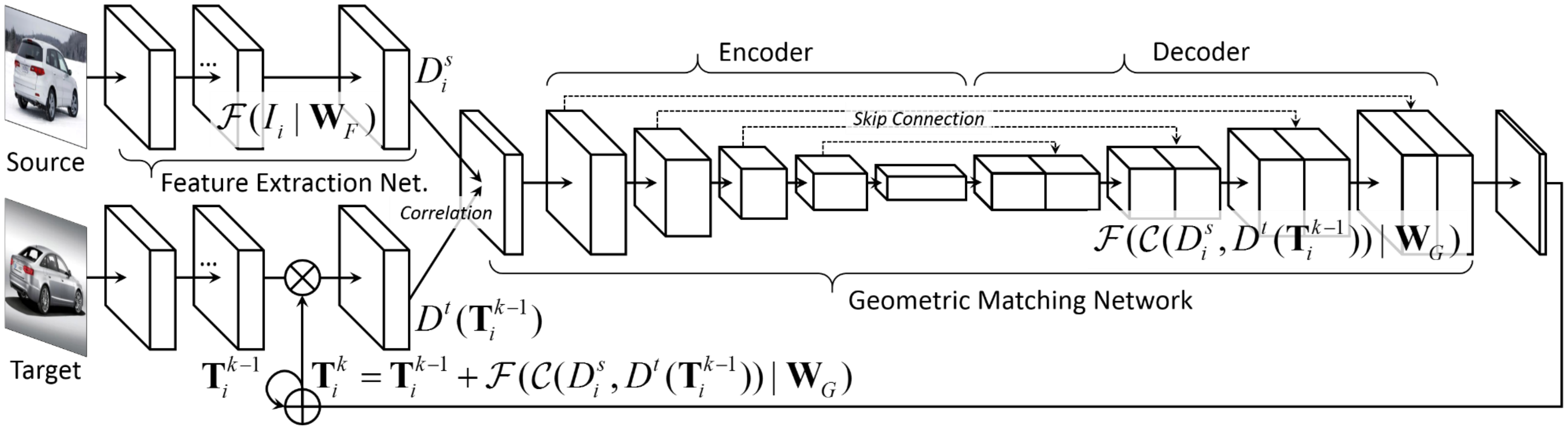
- Constraint correlation volume construction

$$C(D_i^s, D^t(\mathbf{T}_j)) =< D_i^s, D^t(\mathbf{T}_j) >/ \sqrt{< D_i^s, D^t(\mathbf{T}_j) >^2}$$



Recurrent Transformer Networks (RTNs)

Recurrent Geometric Matching Networks



- Recurrent geometric inference

$$\mathbf{T}_i^k - \mathbf{T}_i^{k-1} = F(C(D_i^s, D^t(\mathbf{T}_i^{k-1}))|\mathbf{W}_G)$$

Source Target Iter. 1 Iter. 2 Iter. 3 Iter. 4

Recurrent Transformer Networks (RTNs)

Weakly-supervised Learning

- **Intuition**: matching score between the source D^s at each pixel i and the target $D^t(\mathbf{T}_i)$ should be maximized while keeping the scores of other candidates low

- **Loss Function**:

$$L\left(D_i^s, D^t(\mathbf{T})\right) = -\sum_{j \in M_i} p_j^* \log(p(D_i^s, D^t(\mathbf{T}_j)))$$

where the function $p(D_i^s, D^t(\mathbf{T}_j))$ is a Softmax probability

$$p(D_i^s, D^t(\mathbf{T}_j)) = \frac{\exp(C(D_i^s, D^t(\mathbf{T}_j)))}{\sum_{l \in M_i} \exp(C(D_i^s, D^t(\mathbf{T}_l)))}$$

where p_j^* denotes a class label defined as 1 if $j = i$, 0 otherwise

Experimental Results

Results on the TSS Benchmark

Source images	Target images	SCNet [Han *et al.*, ICCV'17]	GMat. w/Inl. [Rocco *et al.*, CVPR'18]	**RTNs**

Experimental Results

Results on the PF-PASCAL Benchmark


Source images


Target images


SCNet
[Han *et al.*, ICCV'17]


GMat. w/Inl.
[Rocco *et al.*, CVPR'18]


RTNs

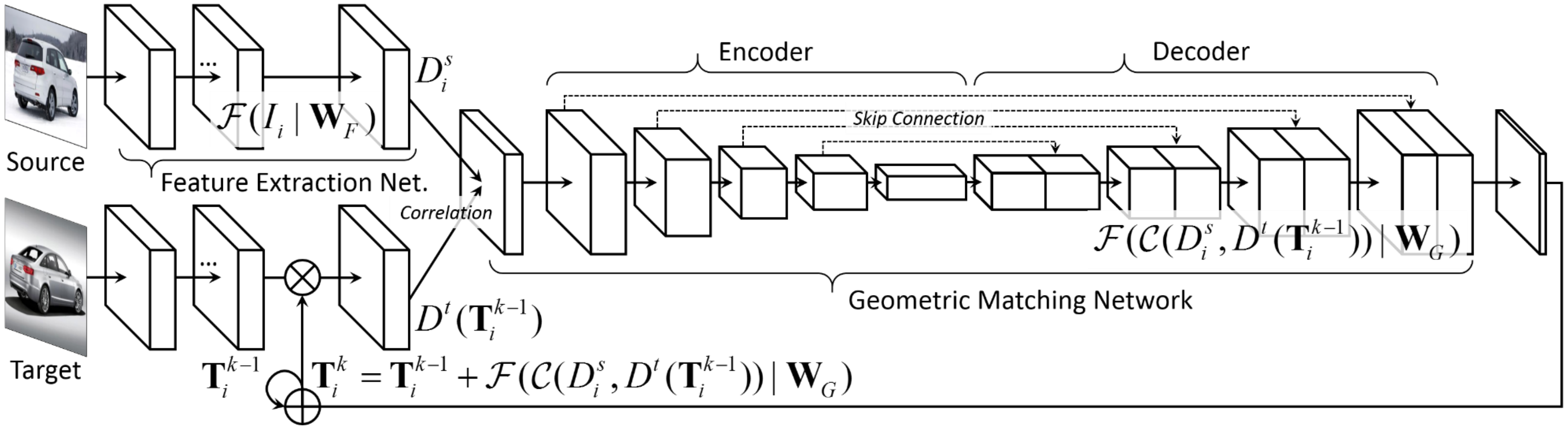
Experimental Results

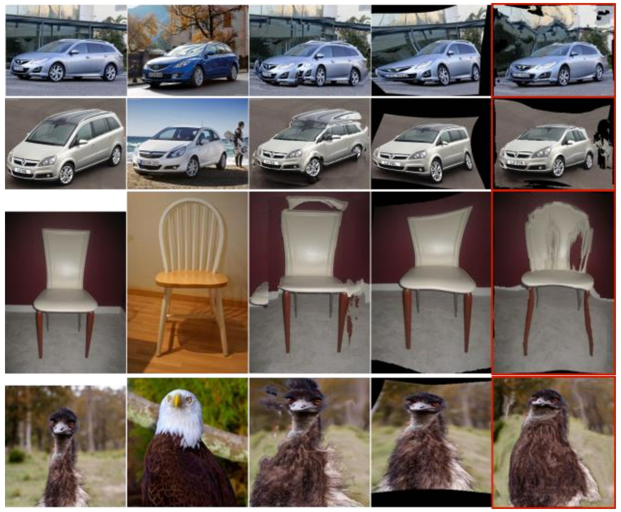
Results on the PF-PASCAL Benchmark

Source images | Target images | SCNet [Han *et al.*, ICCV'17] | GMat. w/Inl. [Rocco *et al.*, CVPR'18] | **RTNs**

Concluding Remarks

- RTNs learn to infer **locally-varying geometric fields** for semantic correspondence in an end-to-end and weakly-supervised fashion
- The key idea is to utilize and iteratively refine **the transformations and convolutional activations through matching** between the image pair
- A technique is presented for **weakly-supervised training** of RTNs





Thank you!

See you at 210 & 230 AB #119

Seungryong Kim, Ph.D.
Digital Image Media Lab.
Yonsei University, Seoul, Korea
Tel: +82-2-2123-2879
E-mail: srkim89@yonsei.ac.kr
Homepage: http://diml.yonsei.ac.kr/~srkim/