JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

THU-AM-125

Correspondence Transformers with Asymmetric Feature Learning and Matching Flow Super-Resolution

Yixuan Sun[1,*], Dongyang Zhao[2], Zhangyue Yin[2], Yiwen Huang[2], Tao Gui[2] and Wenqiang Zhang[1,2], Weifeng Ge[2,†]

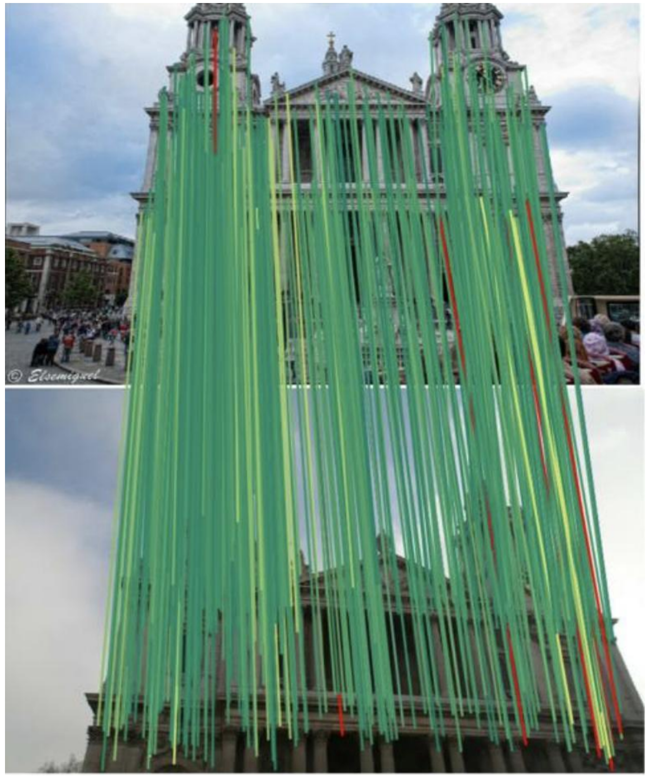
[1]Academy for Engineering and Technology, Fudan University [2]School of Computer Science, Fudan University

{wfge}@fudan.edu.cn

https://github.com/YXSUNMADMAX/ACTR

Introduction

- Existing Image matching task can be divided as Feature Matching, Dense Matching and Semantic Correspondence.
- Semantic Correspondence Aims To Establish Pixel-level Correspondence between Semantically Adjacent Image Pair.



Task of Feature Matching

Task of Dense Matching

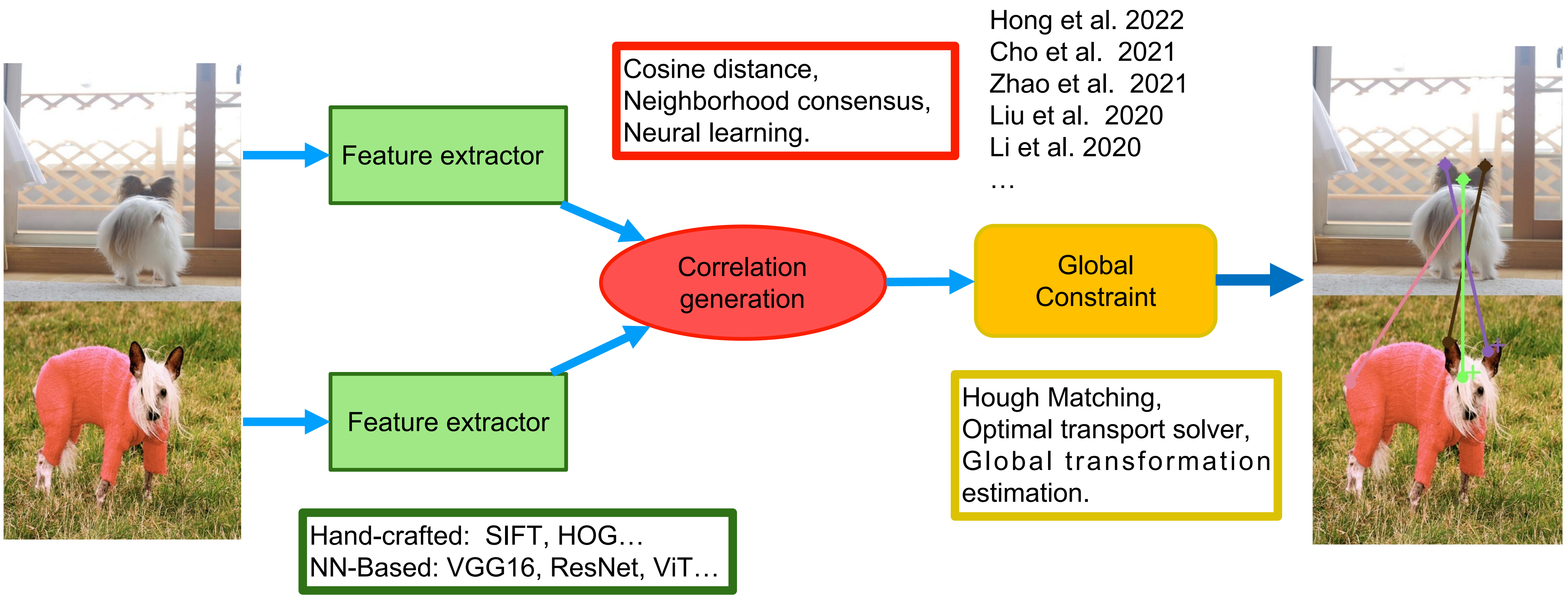
Task of Semantic Correspondence

Challenges

- Large intra-class variation.
- Requires high-quality patch-level representations with aligned semantic spaces.
- Requires matching representation in high resolution.

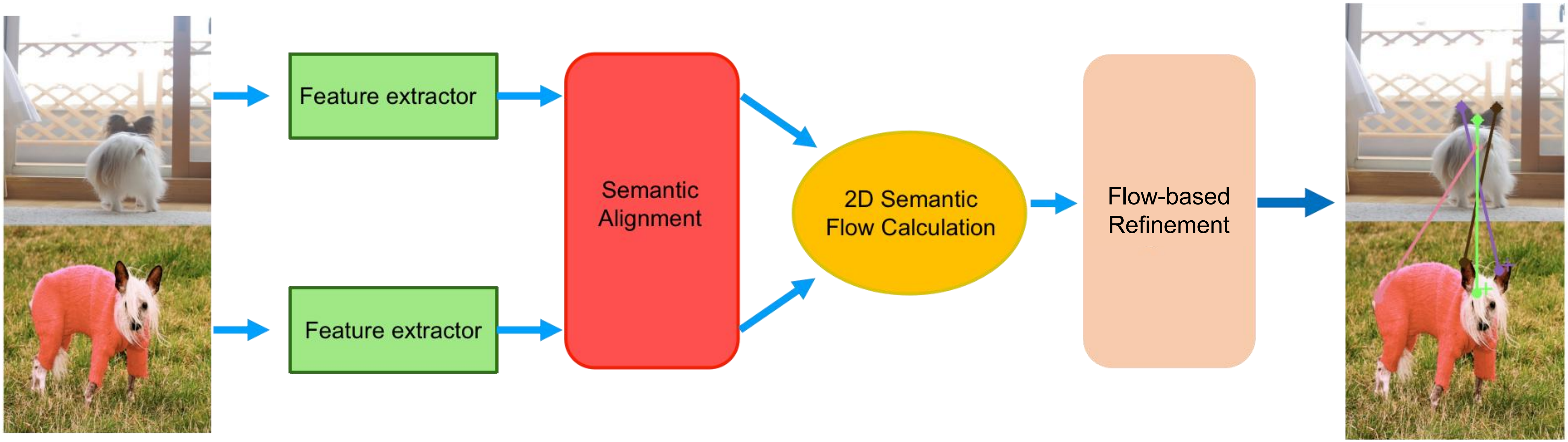
Previous Frameworks

- Siamese Backbone = Shared Semantic Space
- 4D Matrix-based Refinement.



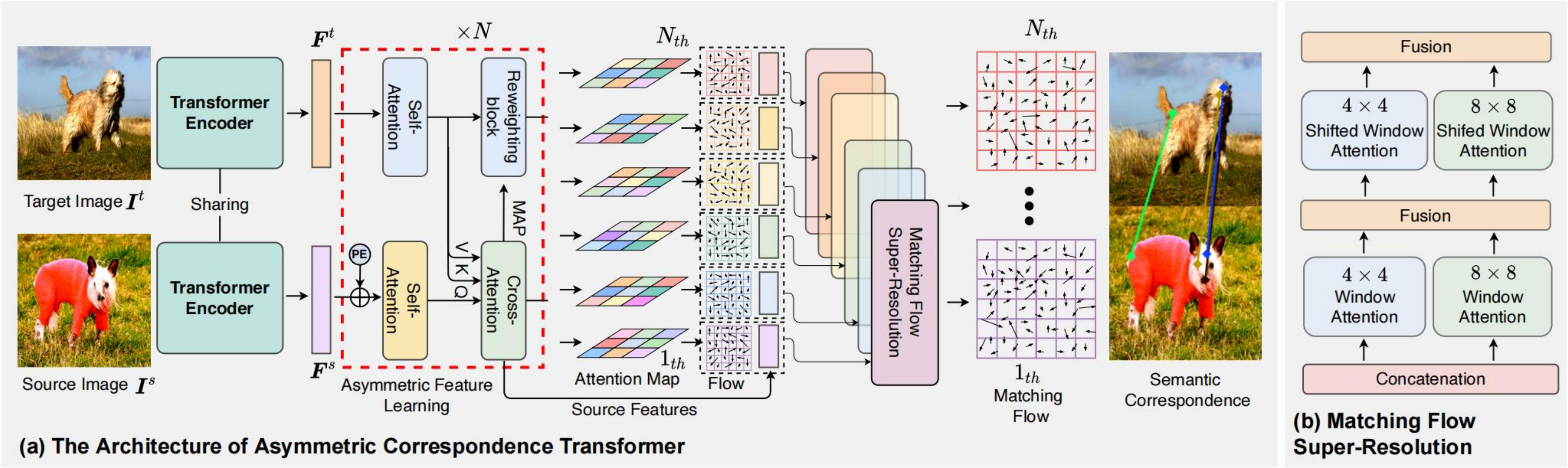
Further Incremental Designs

- Further Semantic Alignment.
- 2D Semantic Flow based Refinement.

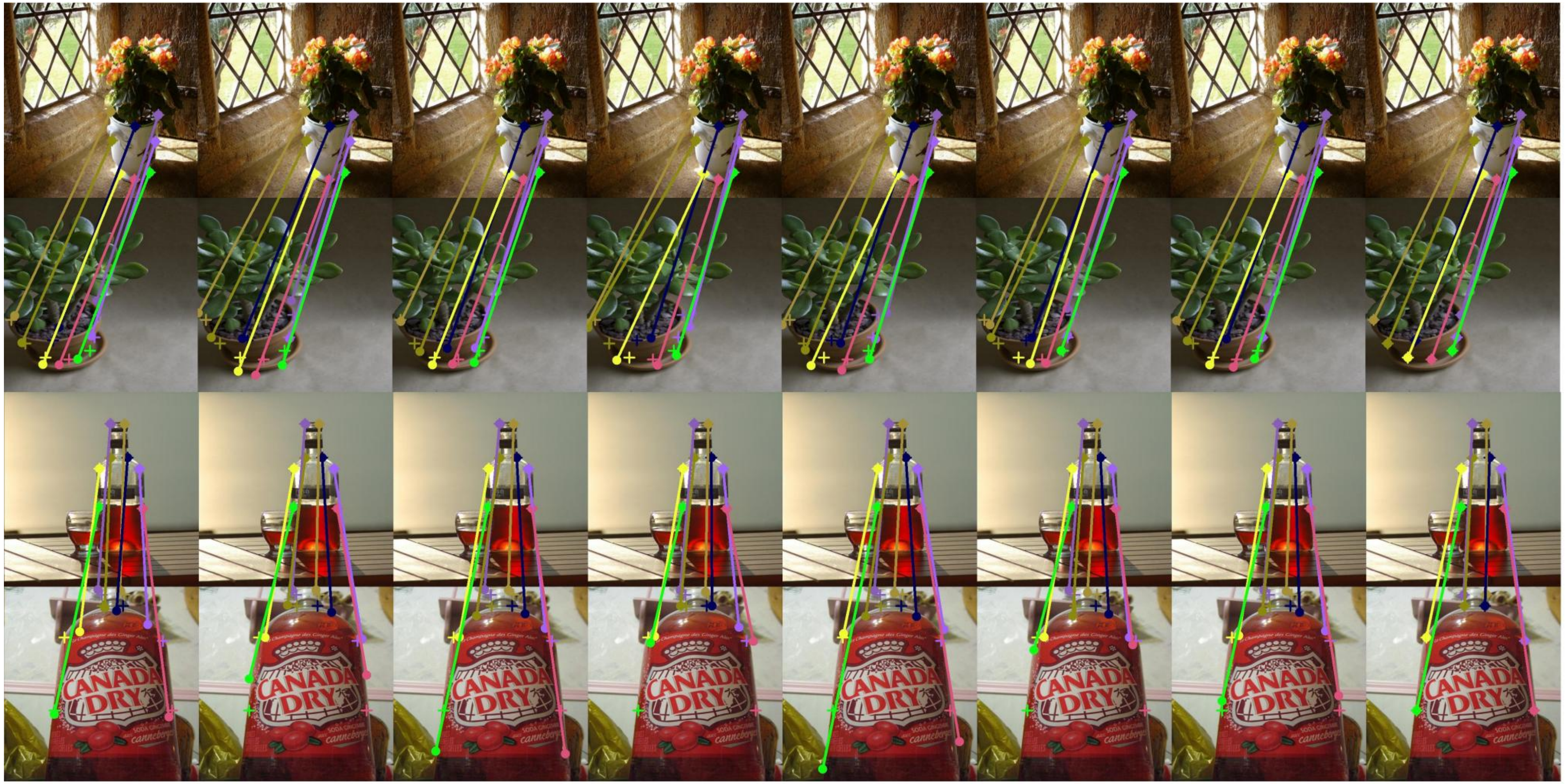


Our Approach

- Asymmetric Alignment Block.
- Multi-Path Semantic Flow Superresolution Block.



The usage of Multi-path Fusion



(a) Path 1 (b) Path 2 (c) Path 3 (d) Path 4 (e) Path 5 (f) Path 6 (g) Fused (h) Ground Truth

Combines the Advantages of Matching Results Under Different Alignment Levels

Results

(a) SCOT (b) CATs (c) MMNet (d) ACTR(ours) (e) GroundTruth

ACTR can clearly distinguish the subtle semantic differences which usually leads to mismatching for previous methods.

Comparison with other methods

(a) Source (b) Target (c) SCOT (d) CATs (e) MMNet (f) ACTR(ours) (g) GroundTruth

Liu et al. 2020; Cho et al. 2021; Zhao et al. 2021;

Evaluation

We set ACTR in 256x256 Resolution as our base model.

Methods	Backbone	Input Resolution	Multi-Scale	SPair-71K α : bbox 0.1	PF-PASCAL α : img 0.05	0.1	0.15
SCOT [30]	ResNet-101	$300 \times 300^*$	✓	35.6	63.1	85.4	92.7
DHPF [35]	ResNet-101	240×240	✓	37.3	75.7	90.7	95
CHM [33]	ResNet-101	256×256	×	46.3	80.1	91.6	94.9
CATs [8]	ResNet-101	256×256	✓	49.9	75.4	92.6	96.4
MMNet-FCN [49]	ResNet-101	224×320	✓	50.4	81.1	91.6	95.9
TransforMatcher [24]	ResNet-101	240×240	✓	53.7	80.8	91.8	-
CATs [8]	iBOT-B	256×256	✓	55.2	77.8	93.1	96.8
TransforMatcher [24]	iBOT-B	240×240	✓	57.9	77.3	93.3	96.6
Baseline	iBOT-B	256×256	×	57.7	78.9	93.2	96.5
ACTR	iBOT-B	256×256	×	**62.1**	**81.2**	**94.0**	**97.0**
VAT [20]	ResNet-101	512×512	✓	54.2	-	92.3	-
VAT [20]	iBOT-B	512×512	✓	59.0	73.0	92.6	**96.7**
Baseline_h	iBOT-B	512×512	×	61.6	79.3	91.6	95.9
ACTR_h	iBOT-B	512×512	×	**65.4**	**82.0**	**93.5**	**96.7**

Yields large Improvements over several benchmarks.

Evaluation

We set ACTR in 256x256 Resolution as our base model.

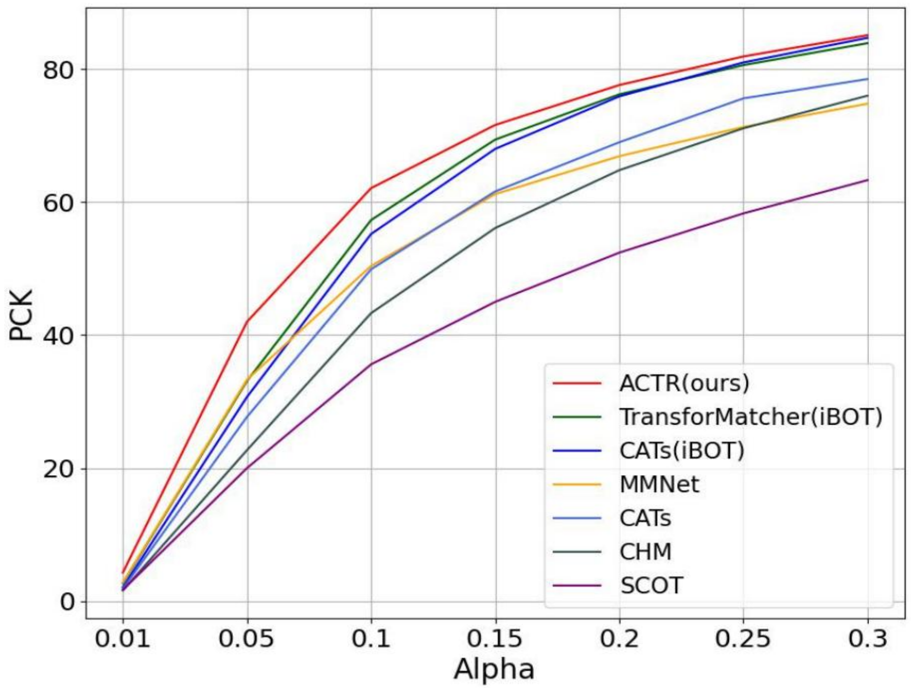
Methods	aero.	bike	bird	boat	bott.	bus	car	cat	chai	cow	dog	hors.	mbik.	pers.	plan.	shee.	trai.	tv	all
SCOT [9]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20	42.9	42.5	31.1	29.8	35	27.7	24.4	48.4	40.8	35.6
DHPF [13]	38.4	23.8	68.3	18.9	42.6	27.9	20.1	61.6	22	46.9	46.1	33.5	27.6	40.1	27.6	28.1	49.5	46.5	37.3
CATs [3]	52	34.7	72.2	34.3	49.9	57.5	43.6	66.5	24.4	63.2	56.5	52	42.6	41.7	43	33.6	72.6	58	49.9
MMNet [16]	55.9	37	65	35.4	50	63.9	45.7	62.8	28.7	65	54.7	51.6	38.5	34.6	41.7	36.3	77.7	62.5	50.4
TransforMatcher [8]	59.2	39.3	73.0	41.2	52.5	**66.3**	55.4	67.1	26.1	67.1	56.6	53.2	45.0	39.9	42.1	35.3	75.2	68.6	53.7
CATs‡ [3]	56.7	41.3	77.8	35.0	54.8	59.8	45.2	69.9	31.4	63.7	57.6	62.5	46.7	49.1	43.2	43.5	76.4	64.1	55.2
TransforMatcher‡ [8]	57.1	47.4	**83.5**	42.3	**56.8**	57.0	55.4	**75.3**	34.5	66.1	64.2	60.2	52.8	55.2	40.5	46.0	75.1	65.8	57.9
ACTR	**65.1**	**48.5**	82.3	**50.4**	55.9	65.3	**63.1**	72.8	**35.8**	**74.1**	**70.3**	**68.9**	**58.6**	**57.1**	**46.8**	**49.5**	**84.4**	**73.3**	**62.1**
VAT [6]	56.5	37.8	73.0	38.7	50.9	58.2	40.8	70.5	20.4	72.6	61.1	57.8	45.6	48.1	52.4	39.7	77.7	**71.4**	54.2
VAT‡	58.6	47.8	83.2	45.6	52.4	67.1	61.4	73.4	30.2	76.5	67.7	66.9	48.0	53.3	46.6	44.3	84.6	60.7	59.0
ACTR$_h$	**64.9**	**54.8**	**87.6**	**49.2**	**55.7**	**74.4**	**66.5**	**80.7**	**35.3**	**82.1**	**75.2**	**71.9**	**54.0**	**62.4**	**54.9**	**53.5**	**88.7**	71.0	**65.4**

Yields large Improvements on a challenging dataset.
Reach best result on 14/18 sub-classes.

Evaluation

We set ACTR in 256x256 Resolution as our base model.

Methods	PF-WILLOW α : bbox 0.05	PF-WILLOW α : bbox 0.1	PF-WILLOW α : bkp 0.05	PF-WILLOW α : bkp 0.1
DHPF [35]	49.5	77.6	-	71.0
CHM [33]	52.7	79.4	-	69.6
CATs [8]	50.3	79.2	40.7	69.0
SCOT [30]	-	-	47.8	76.0
TransforMatcher [24]	-	65.3	-	76.0
CATs‡	59.4	86.3	51.1	79.5
TransforMatcher‡	57.0	84.3	48.8	78.3
ACTR	**60.3**	**87.2**	**52.6**	**79.9**



Yields better generalizability when testing on PF—WILLOW.

Ablation Results

We set ACTR in 256x256 Resolution as our base model.

Methods	SPair-71K $\alpha_{bbox} = 0.1$
ACTR	62.1
w/o source branch positional encoding	60.4 (1.7↓)
w/o target branch token reweighting	60.7 (1.4↓)
w/o asymmetric cross attention module	60.1 (2.0↓)
w/o multi-path super-resolution	61.0 (1.1↓)
w/o dual window flow refinement	60.6 (1.5↓)
w/o flow super-resolution module	59.0 (3.1↓)
Baseline	57.7 (4.4↓)

Design of asymmetric alignment and multi—path super—resolution can help to improve the accuracy in semantic correspondence

Thank You

Academy of Engineering & Technology, Fudan University, Shanghai, China
School of Computer Science, Fudan University, Shanghai, China
{wfge}@fudan.edu.cn