MATCHA : Towards Matching Anything

Xue, Fei, et al. "MATCHA: Towards matching anything." *Proceedings of the* ***Computer Vision and Pattern Recognition Conference***. **2025**.

발표자 : 홍권

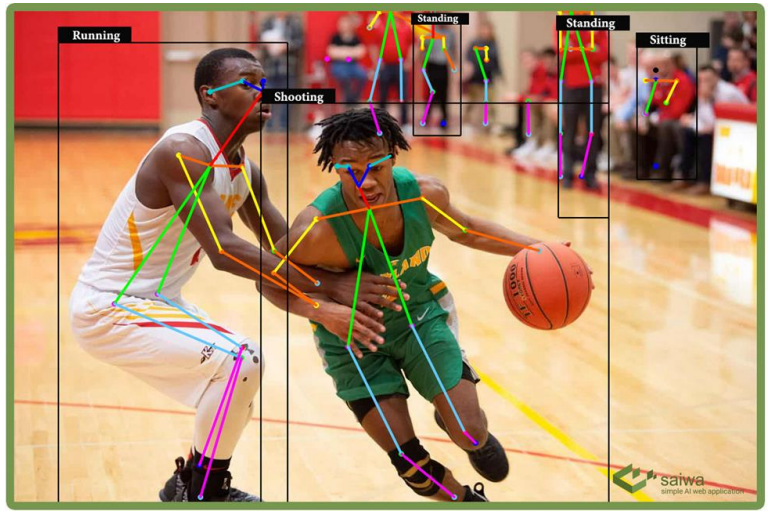
Contents

Contents

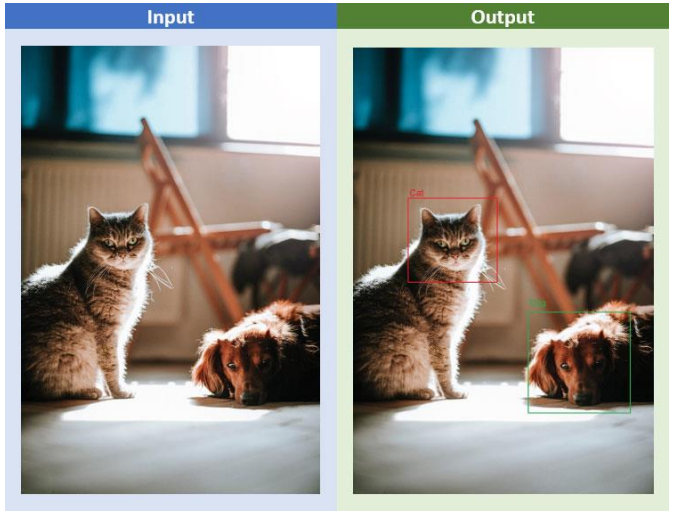
Introduction

"In computer vision, there is only one problem :
Correspondence, Correspondence, Correspondence ... "

- 컴퓨터 비전의 가장 근본적인 문제는 이미지들 간의 **대응 관계**를 설정하는 것
  - Mapping, Localization, Image Editing, Object Pose Estimation, Point Tracking ,,,



Human Pose Estimation



Localization

Introduction

컴퓨터 비전의 task들은 **세 가지 특징**으로 구분 가능

- **Geometric** : 현실 세계의 3D에 해당하는 정보를 정적 이미지의 2D point로 식별 → **다양한 조명 조건, 시점 변화**
- **Semantic** : 동일 범주의 서로 다른 객체들 사이에서 유사한 객체 부위 연결 → **서로 다른 객체 간 고차원적인 추상화를 요구**
- **Temporal** : 동일 객체의 점을 영상 프레임에 매칭 → **가려짐, 변형, 동적 요소 등의 다양한 처리 능력 요구**

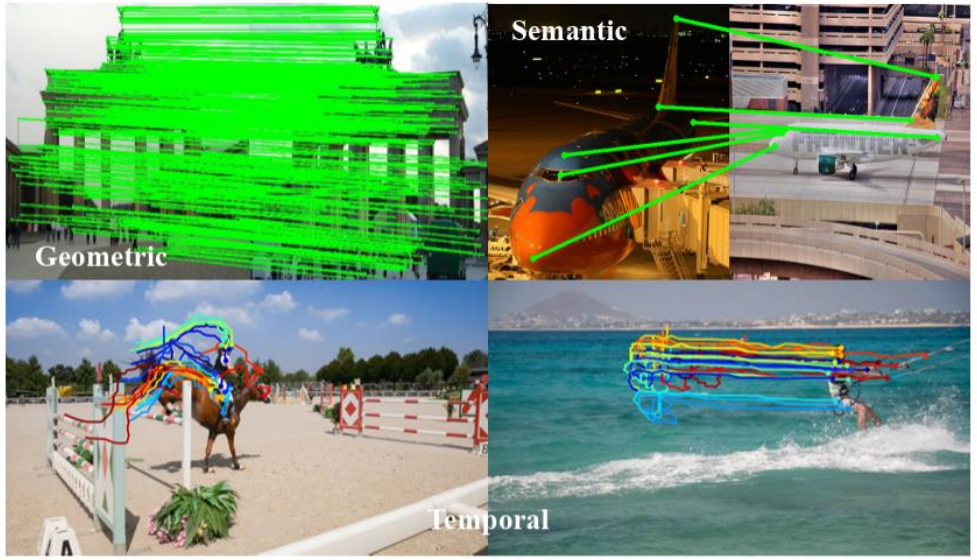


Figure 1. **MATCHA for *matching anything*.** We visualize geometric, semantic and temporal correspondences established by MATCHA, using a single feature descriptor.

Introduction

- 과거 연구들은 Geometric, Semantic, Temporal 각 **task 별 특화 모델**을 사용해 문제를 해결
- 그러나, 사람은 다양한 task를 유연하게 처리 → **task마다 별도의 모델이 필요할까 ??**

MATCHA : Foundation Model for matching Anything

- **DIFT**로부터 아이디어 채용 [Emergent Correspondence for Image Diffusion]
  - Diffusion 네트워크에 암묵적으로 존재하는 지식을 활용해 이미지 특징으로 추출
- DIFT와 달리, 단일 descriptor 를 활용
- **DINOv2** Foundation Model의 지식 + 지도 학습 → **정확도**와 **일반화 성능** 향상
  - Foundation Model 대비, 지도 학습 데이터 규모 제한적
  - 또한, Semantic 및 Temporal 데이터는 고비용

- **DINOv2**

Contents

Related Works

- **Geometric Matching**
  - 동일한 장면을 촬영한 두 이미지 사이에 **물리적으로 대응하는 지점을 찾는 것**
  - 지역을 **검출(Detecting), 기술(Describing), 매칭(Matching)** 함으로써 설정
  - 기존 연구에서는 **기하학적(Geometric)** 특징에 초점을 맞추느라 의미론적(Semantic)한 특징을 살리지를 못함

- **Semantic Matching**
  - 동일한 카테고리에 속하는 다른 객체들 사이에서 의미적으로 유사한 의미를 가진 지점을 매칭 하는 것
  - 기존 연구들은 방대한 양의 데이터로 학습되어 **풍부한 Semantic 지식을 갖는 Foundation 모델을 활용**

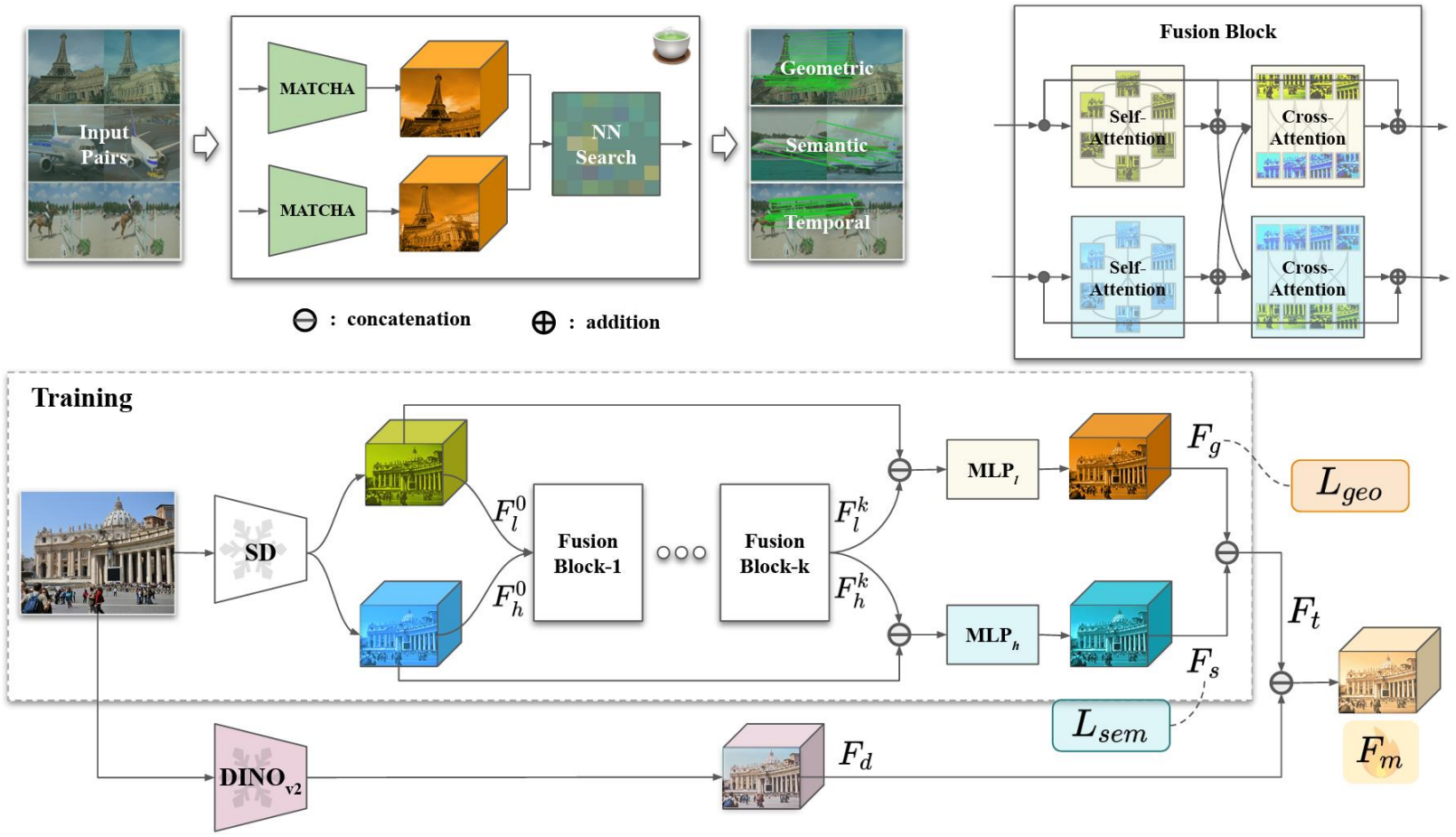
- **Temporal Matching**
  - 시간의 흐름이 있는 데이터에서, **동일 객체의 동일 지점**을 연결하는 것
  - Temporal 매칭이 종합적으로 가장 어려운 문제

Contents

Preliminaries

- **DIFT**
  - Diffusion 모델로부터 feautre를 추출하는 연구에서 영감
  - 이미지 생성을 위해 학습된 **Diffusion 모델은 암시적으로 Correspondence를 학습**
  - **특정 Layer와 Timestep에서 feature를 추출**함으로써 Geometric, Semantic, Temporal 매칭 task에 효과적
  - $I \in R^{H \times W \times 3}$의 RGB 이미지, DIFT는 사전 학습된 Stable Diffusion으로부터 다음을 추출
    - ❖ $F_h \in R^{\frac{H}{16} \times \frac{W}{16} \times 3}$ (Semantic descriptor)
    - ❖ $F_l \in R^{\frac{H}{8} \times \frac{W}{8} \times 640}$ (Geometric descriptor)
- **DINOv2**
  - 수 백만 장의 이미지로 학습된 **Self-Supervised Foundation Model**
  - DINOv2는 개별 객체에 대해 극단적인 시점 변화 및 크기 변화를 유연하게 처리, Temporal 매칭 task에 뛰어난 성능을 보임
  - 그러나, Spatial Detail 부족 → Geometric 매칭 성능 한계

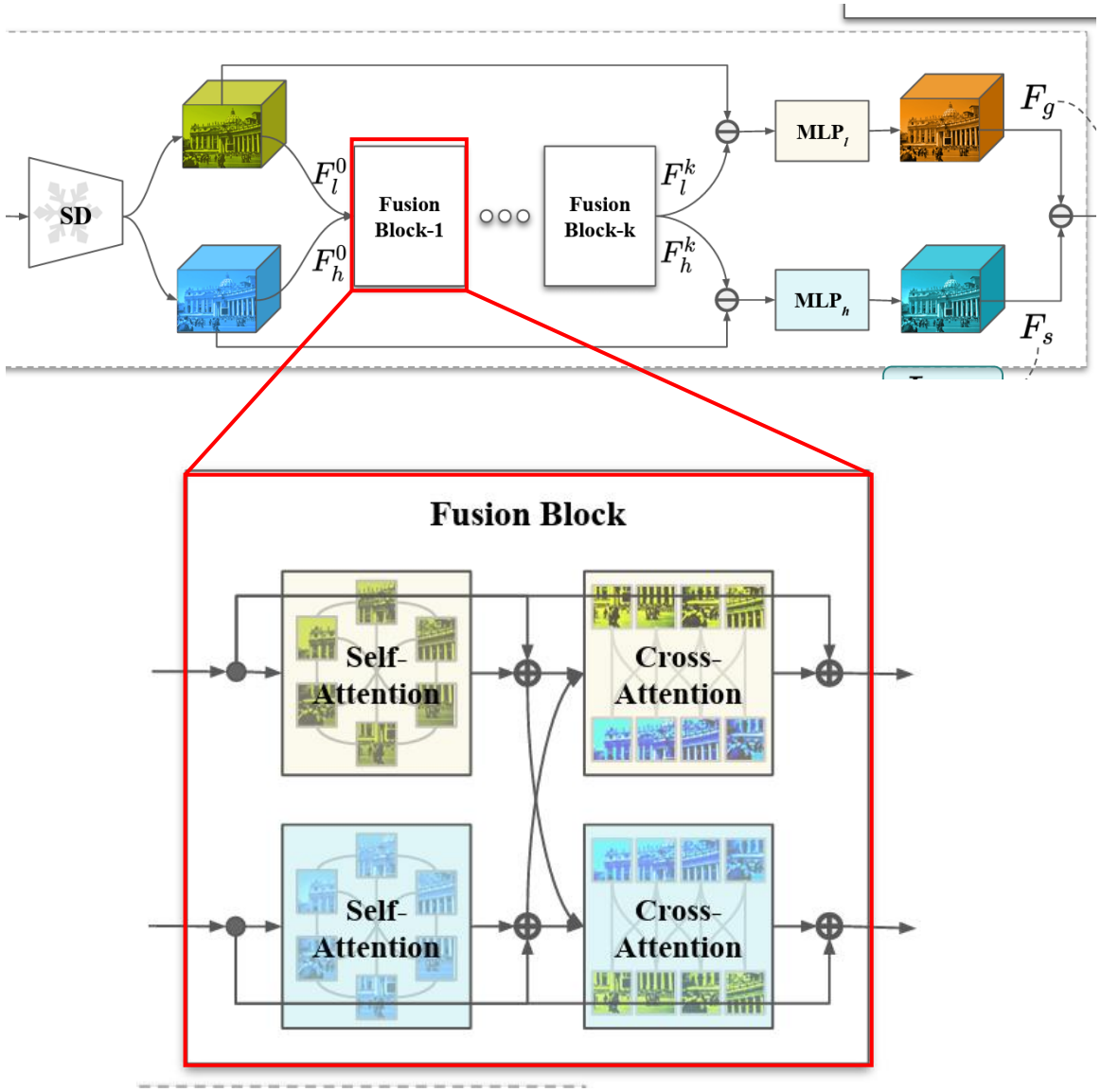
Architecture



1. **DINOv2**와 **DIFT**를 활용하여 특징 추출
   - Semantic descriptor : F_h, F_d
   - Geometric descriptor : F_l
2. **DIFT**의 F_h와 F_l이 서로 보완 되도록 강화
   - **Dynamic Fusion Model**
   - 지도 학습
3. 강화된 F_h와 F_l을 **DINOv2**의 F_d를 병합하여 최종 descriptor F_m 생성

- **입력 :** RGB 이미지
- **출력 :** $F_m \in R^{\frac{H}{8} \times \frac{W}{8} \times D_m}$

Dynamic Feature Fusion



- 동적 융합을 위해 트랜스포머 구조 채택
  - Semantic과 Geometric descriptor로부터 정보를 동적 수집

1. F_h와 F_l 을 패치 크기 p로 패치화
2. Linear Layer를 통해 공통의 feature 차원으로 투영
3. 위 과정을 통해 F_h^0, F_l^0 (Semantic, Geometric) $\in R^{N \times D^h}$ 생성, N은 패치화 후 변화된 총 feature 수
4. Dynamic Fusion 블록에 F_h^0, F_l^0 통과
   - Dynamic Fusion 블록은 k개의 Self-Attention 및 Cross-Attention 블록으로 구성
   - Dynamic Fusion 안의 업데이트 과정
     1) $F_h^i = F_h^{i-1} + self_h^i(F_h^{i-1})$
     2) $F_l^i = F_l^{i-1} + self_l^i(F_l^{i-1})$
     3) $F_h^i = F_h^i + cross_h^i(F_h^i, F_l^i)$
     4) $F_l^i = F_l^i + cross_l^i(F_h^i, F_l^i)$
5. 원본 입력 F_h^0, F_l^0 과 Dynamic Fusion 블록 출력 F_h^k, F_l^k를 연결 후, MLP 통과
   1) $F_s = MLP_h([F_h^0 \| F_h^k])$
   2) $F_g = MLP_h([F_l^0 \| F_l^k])$

Feature Merging & Supervision

- **Feature Merging**
  - Dynamic Feature Fusion에서 **생성된 feature (F_s, F_g)를 연결하여 매끄러운 feature를 만들어내는 과정**
  - $F_s + F_g$ 를 연결하여 F_t 생성, F_t는 **Geometric 정보와 Semantic 정보**를 모두 포함
  - F_t와 **DINOv2**에서 얻은 F_d를 연결하여 **최종 feature F_m 구성**

- **Supervision**
  - 최종 feature인 F_m 에 직접 지도 학습 적용 X, **오로지 Dynamic Fusion 과정에만 지도 학습** 적용
  - **Why ?**
    - Foundation 모델 학습 데이터 수 대비, 지도 학습 데이터 수의 불균형 큼
    - Semantic 및 Temporal 데이터는 고비용
  - Semantic feature Fine-tuning : **Contrastive Loss + Dense Semantic Flow Loss**
  - Geometric feature Fine-tuning : **Dual Softmax Loss**

Contents

Experimetns

- **Semantic Matching**

Method	SM. Sup.	SPair-71k [40] PCK$_{@0.01/0.05/0.1}$(↑)	PF-Pascal [18] PCK$_{@0.05/0.1/0.15}$(↑)	PF-Willow [17]
DINOv2 [44]	✗	6.3 / 38.4 / 53.9	63.0 / 79.2 / 85.1	43.8 / 75.4 / 86.1
*DIFT [60]	✗	7.2 / 39.7 / 52.9	66.0 / 81.1 / 87.2	58.1 / 81.2 / -
DIFT	✗	3.1 / 37.9 / 54.3	58.7 / 81.8 / 87.8	55.7 / 85.1 / 92.9
USC [20]	✗	- / 28.9 / 45.4	-	53.0 / 84.3 / -
SD+DINO [74]	✗	7.9 / 44.7 / 59.9	71.5 / 85.8 / 90.6	-
†GeoASM [73]	✗	9.9 / 49.1 / 65.4	74.0 / 86.2 / 90.7	-
DHF [38]	✓	8.7 / 50.2 / 64.9	78.0 / 90.4 / 94.1	-
*SCorrSAN [21]	✓	3.6 / 36.3 / 55.3	81.5 / 93.3 / 96.6	54.1 / 80.0 / 89.8
*CATs++ [5]	✓	4.3 / 40.7 / 59.8	84.9 / 93.8 / 96.8	56.7 / - / 81.2
*SD4Match [31]	✓	- / 59.5 / 75.5	84.4 / 95.2 / 97.5	56.7 / 80.9 / 91.6
*SD+DINO [74]	✓	9.6 / 57.7 / 74.6	80.9 / 93.6 / 96.9	-
*†GeoASM [73]	✓	**22.0 / 75.3 / 85.6**	**85.9 / 95.7 / 98.0**	-
MATCHA-Light	✓	10.4 / 65.5 / 78.9	82.3 / 93.5 / 96.6	69.0 / 90.1 / 96.2
MATCHA	✓	12.2 / 67.1 / 79.6	79.5 / 93.0 / 96.8	**70.2 / 91.3 / 97.0**

Table 1. **Evaluation on Semantic Matching.** We report PCK under different thresholds. * denotes methods with dataset-specific models and † denotes semantic masks being required. Red indicates methods using image pairs as inputs. Both results of DIFT from its original paper [60] (*DIFT) and our implementation (DIFT) are included.

- **Geometric Matching**

Method	GM Sup.	MegaDepth [32]	ScanNet [9] AUC$_{@5/10/20}$(↑)	Aachen [56]
Croco.E [67] + SP	✗	8.0 / 14.7 / 24.2	1.8 / 4.2 / 8.4	11.4 / 18.2 / 26.3
DINOv2 [44] + SP	✗	24.6 / 37.4 / 50.9	2.3 / 5.9 / 12.3	17.2 / 26.1 / 36.4
DIFT [60] + SP	✗	49.7 / 62.8 / 72.8	9.3 / 18.7 / 29.4	43.7 / 53.1 / 61.3
SP [10]	✓	47.2 / 60.0 / 69.9	6.8 / 14.9 / 24.7	41.6 / 50.2 / 58.1
XFeat [46]	✓	45.4 / 58.9 / 69.3	12.3 / 25.9 / 40.6	36.1 / 45.9 / 55.1
DISK [63]	✓	55.4 / 67.7 / 76.7	6.8 / 14.9 / 24.7	48.9 / 57.5 / 64.6
R2D2 [48]	✓	39.6 / 54.3 / 66.2	5.4 / 11.3 / 19.3	27.6 / 36.4 / 44.1
D2Net [13]	✓	32.5 / 47.7 / 61.4	10.6 / 22.9 / 37.3	30.3 / 41.8 / 52.5
MASt3R.E [30] + SP	✓	37.8 / 51.6 / 63.6	7.4 / 16.8 / 28.5	31.2 / 41.3 / 51.3
MATCHA-Light + SP	✓	**57.1 / 70.9 / 81.2**	**13.0 / 26.6 / 41.8**	51.4 / 60.1 / 67.1
MATCHA + SP	✓	55.8 / 69.3 / 80.0	12.7 / 26.1 / 40.8	**51.7 / 61.0 / 68.5**

Table 2. **Evaluation on Relative Pose Estimation.** We report the AUC values at error thresholds of $5°/10°/20°$ on all datasets.

Experimetns

- **Temporal Matching**

Figure 5. **Visualization of temporal matches on TapVID-Davis [12].** Here we visualize several challenging cases for exstablishing temporal correspondences, where MATCHA generally achieves the best performance in handling extreme scale and viewpoint changes, as well as scenes with multiple similar instances. (DIFT* is the adapted DIFT where we use its concatenated semantic and geometric feature for temporal matching for better performance.)

- **Geometric Matching**

Method	Single Desc	Corres. Sup.	*Geometric* Aachen $AUC_{@5/10/20}(\uparrow)$	Avg(↑)	*Semantic* PF-Willow $PCK_{@0.05/0.1/0.15}(\uparrow)$	Avg(↑)	*Temporal* TapVid-Davis $PCK_{@0.05/0.1/0.15}(\uparrow)$	Avg(↑)	Average Score(↑)
DISK [63]	✓	GM	48.9 / 57.5 / 64.6	57.0	10.2 / 17.0 / 23.1	16.8	57.0 / 61.7 / 65.0	61.2	45.0
XFeat [46]	✓	GM	36.1 / 45.9 / 55.1	45.7	25.7 / 40.0 / 48.8	38.2	63.3 / 71.4 / 77.1	70.6	51.5
MASt3R.E [30]	✓	GM	31.2 / 41.3 / 51.3	41.3	24.0 / 42.1 / 54.7	40.3	75.2 / 83.8 / 87.9	82.3	54.6
DIFT [60]	✗	✗	43.7 / 53.1 / 61.3	52.7	55.7 / 85.1 / 92.9	77.9	79.7 / 86.7 / 90.5	85.6	72.1
MATCHA-Light	✗	GM+SM	51.4 / 60.1 / 67.1	<u>59.5</u>	69.0 / 90.6 / 96.2	<u>85.3</u>	78.7 / 86.3 / 90.2	85.1	<u>76.6</u>
DINOv2 [44]	✓	✗	17.2 / 26.1 / 36.4	26.6	43.8 / 75.4 / 86.1	68.4	83.2 / 89.7 / 92.0	88.3	61.1
DIFT.Uni +DINOv2	✓	✗	41.9 / 51.3 / 60.0	51.1	58.7 / 82.9 / 90.7	77.4	86.4 / 91.6 / 93.5	<u>90.5</u>	73.0
MATCHA	✓	GM+SM	**51.7 / 61.0 / 68.5**	**60.4**	**70.2 / 91.3 / 97.0**	**86.2**	**87.8 / 93.5 / 95.5**	**92.3**	**79.6**

Table 4. **Towards Matching Anything with A Unified Feature.** We compare ourselves to various feature models across geometric, semantic and temporal matching and compute the ranking of each method for each task and averaged over tasks. We show that MATCHA is able to achieve the topk averaged ranking among all types of methods using a single feature for matching anything.