The Inherent Ability to Find Semantic Correspondences in Diffusion Models

DMQA Open Seminar (2024. 08. 02)

Data Mining & Quality Analytics Lab.

박태남

발표자 소개

❖ **박태남 (Taenam Park)**

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2023.03 ~ Present)

❖ **Research Interest**

- Diffusion Models
- Image Generation & Synthesis

❖ **Contact**

- taenampark@korea.ac.kr

Contents

❖ **Introduction**

❖ **Methods**

- ✓ Unsupervised Semantic Correspondence Using Stable Diffusion
- ✓ Emergent Correspondence from Image Diffusion
- ✓ Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ **Conclusion**

Introduction

❖ Visual Correspondence

- 컴퓨터 비전 분야에서 기본적인 문제로 다양한 응용 분야에서 활용

1. Semantic Correspondence: 유사한 의미를 공유하는 서로 다른 객체의 픽셀

Semantic Correspondence

Image editing

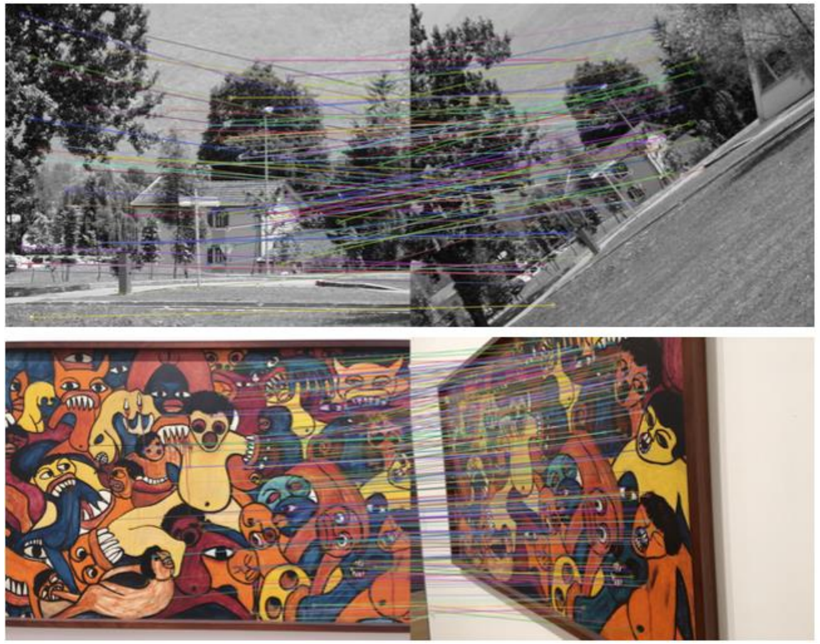
Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.
https://www.cs.cornell.edu/~snavely/bundler

Introduction

❖ Visual Correspondence

- 컴퓨터 비전 분야에서 기본적인 문제로 다양한 응용 분야에서 활용

2. <u>Geometric</u> Correspondence: 다른 시점에서 캡처된 동일한 객체의 픽셀



Geometric Correspondence

3D reconstruction

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.
https://neural-congealing.github.io

Introduction

❖ Visual Correspondence

- 컴퓨터 비전 분야에서 기본적인 문제로 다양한 응용 분야에서 활용

3. Temporal Correspondence: 시간이 지남에 따라 변형될 수 있는 비디오 내 동일한 객체의 픽셀

Temporal Correspondence

Video segmentation

https://co-tracker.github.io
https://medium.com/@eddiesmo/video-object-segmentation-the-basics-758e77321914

Introduction

❖ **Why are diffusion models relevant to finding correspondences?**

- Diffusion models은 이미지 생성 및 합성 분야에서 뛰어난 결과들을 보여주며 주목받고 있음

In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.

Text-to-image diffusion models
(DALLE-3)

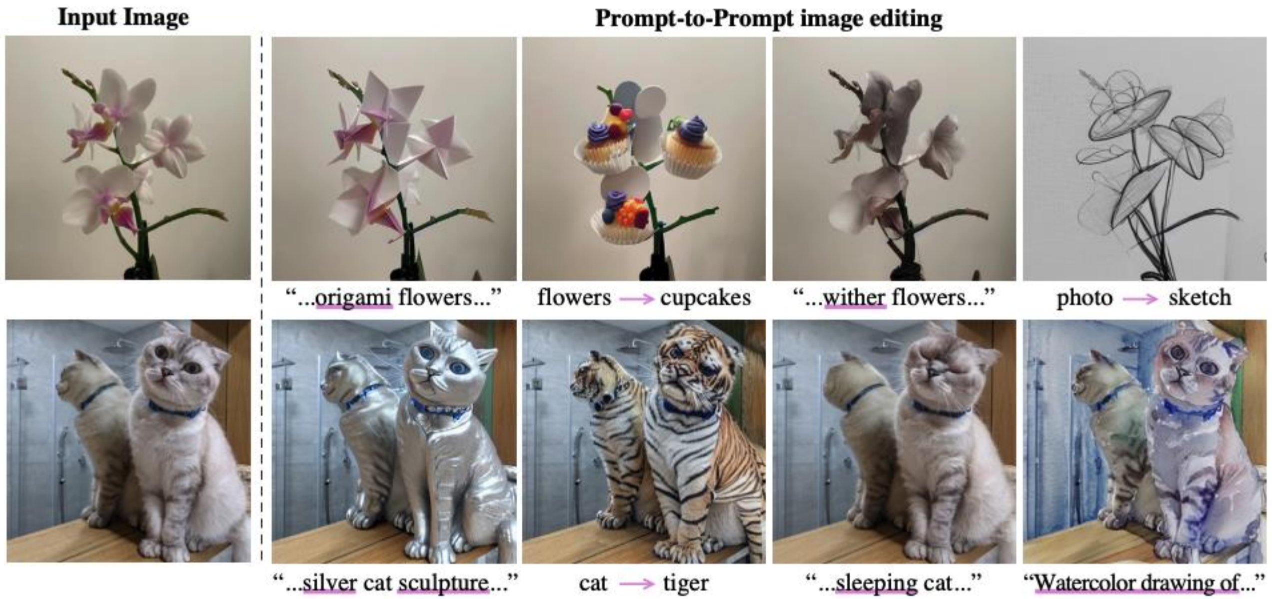


Image-to-image diffusion models
(Null-text Inversion)

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., ... & Ramesh, A. (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, *2*, 3.
Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6038-6047).

Introduction

❖ **Why are diffusion models relevant to finding correspondences?**

- Diffusion models은 이미지 생성 및 합성 분야에서 뛰어난 결과들을 보여주며 주목받고 있음
- 모델은 **생성해야 할 객체의 의미론적 내용을 이해**하고 있거나, **두 범주 간의 대응에 대해 암묵적으로 추론**해야 함

Text-to-image diffusion models
(DALLE-3)

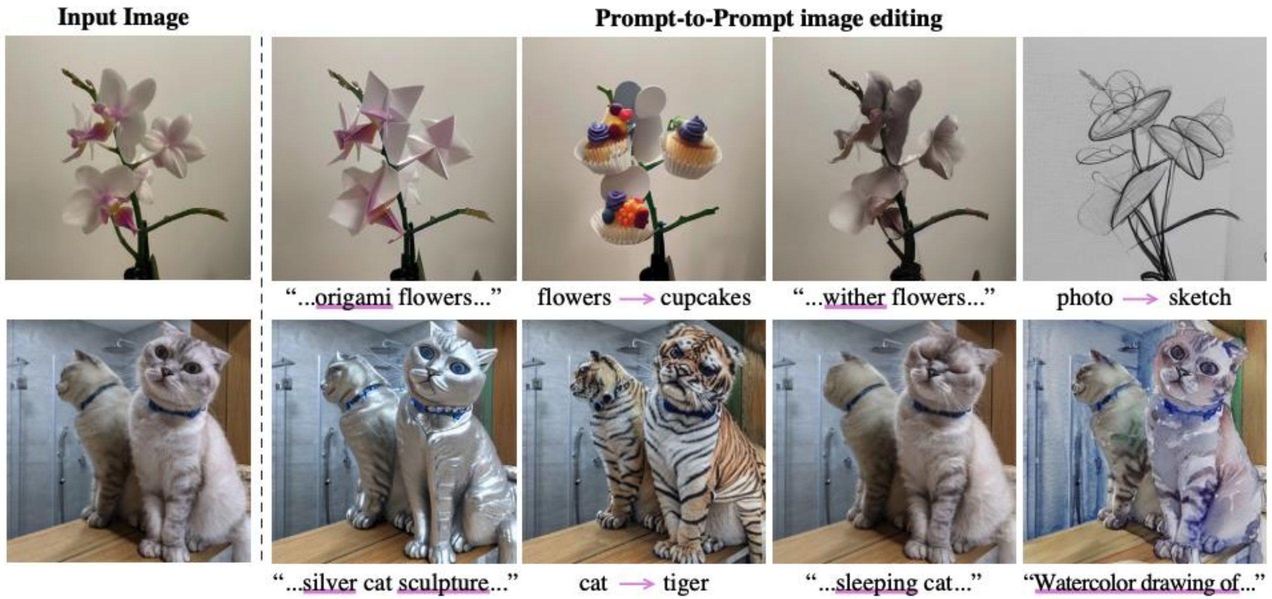
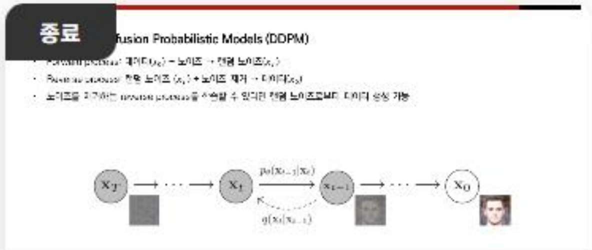


Image-to-image diffusion models
(Null-text Inversion)

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., ... & Ramesh, A. (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, *2*, 3.
Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6038-6047).

Introduction

❖ DMQA Seminar (Diffusion Models)



Score-based Generative Models and Diffu

발표자: 조한샘

2022년 2월 11일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Improving Sampling Speed of Diffusion M

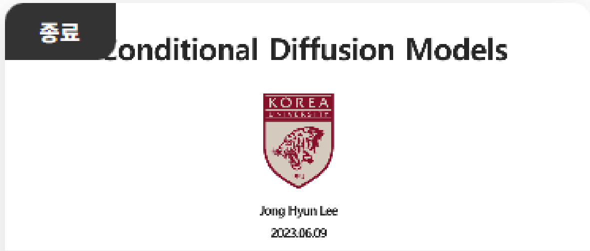
발표자: 조한샘

2023년 2월 10일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶



Conditional Diffusion Models

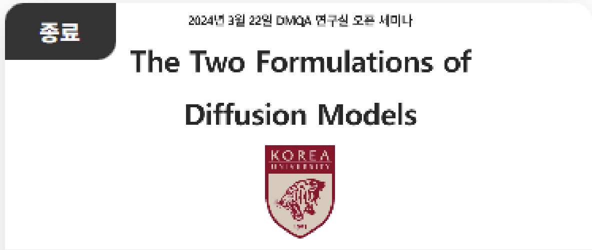
발표자: 이종현

2023년 6월 16일

오전 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶



The Two Formulations of Diffusion Model

발표자: 이종현

2024년 3월 22일

오후 3시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Methods

Text-to-image diffusion models	Image-to-image diffusion models	Image-to-image diffusion models
Unsupervised Semantic Correspondence Using Stable Diffusion	Emergent Correspondence from Image Diffusion	Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

Methods

Text-to-image diffusion models

Unsupervised
Semantic Correspondence
Using Stable Diffusion

Image-to-image diffusion models

Emergent Correspondence
from Image Diffusion

Image-to-image diffusion models

Diffusion Hyperfeatures:
Searching Through
Time and Space
for Semantic Correspondence

Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Unsupervised Semantic Correspondence Using Stable Diffusion(NeurIPS, 2023)

- **Motivation:** Text-to-image diffusion models은 생성해야 할 객체의 의미론적 내용(semantics)을 이해하고 있지 않을까?

’A propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese.’

Text-to-image diffusion model
(SDXL)

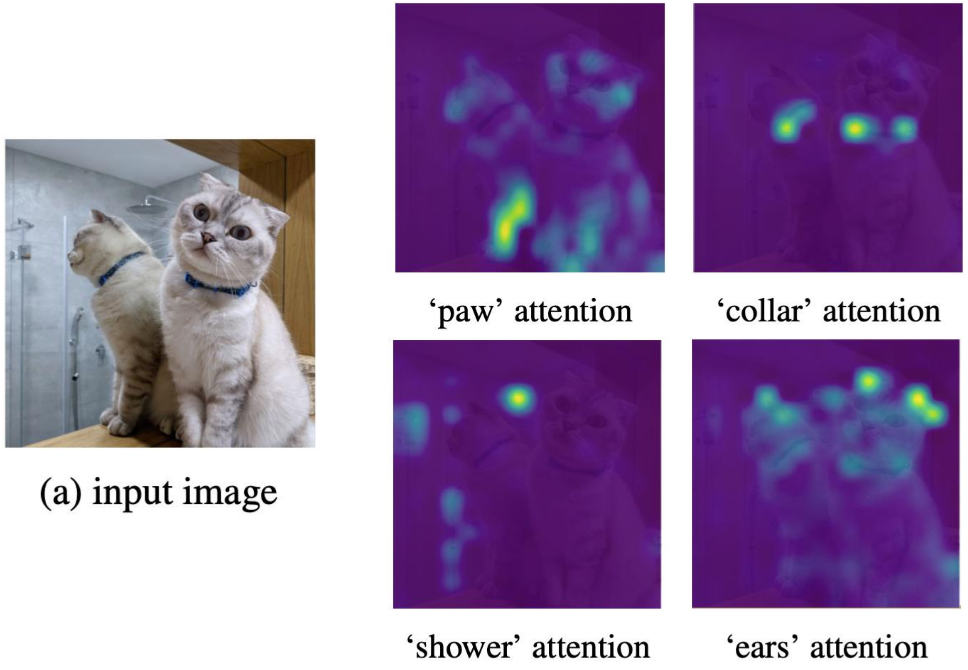
Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Unsupervised Semantic Correspondence Using Stable Diffusion(NeurIPS, 2023)

- **Motivation:** Text-to-image diffusion models은 생성해야 할 객체의 의미론적 내용(semantics)을 이해하고 있지 않을까?



Cross-Attention Map

Attention maps은 prompt의 semantic에 반응함

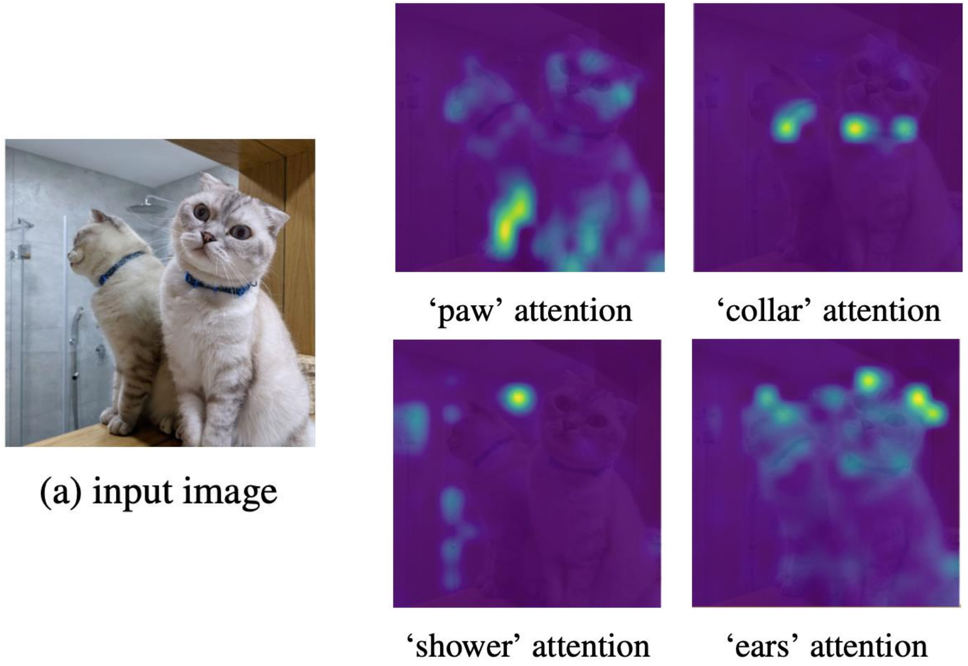
Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Unsupervised Semantic Correspondence Using Stable Diffusion(NeurIPS, 2023)

- **Motivation:** Text-to-image diffusion models은 생성해야 할 객체의 의미론적 내용(semantics)을 이해하고 있지 않을까?



Cross-Attention Map

Attention maps은 prompt의 semantic에 반응함

→ 특정 이미지 위치에 해당하는 프롬프트를 식별한다면, 이미지에서 의미적으로 유사한 이미지 위치 대응 가능

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

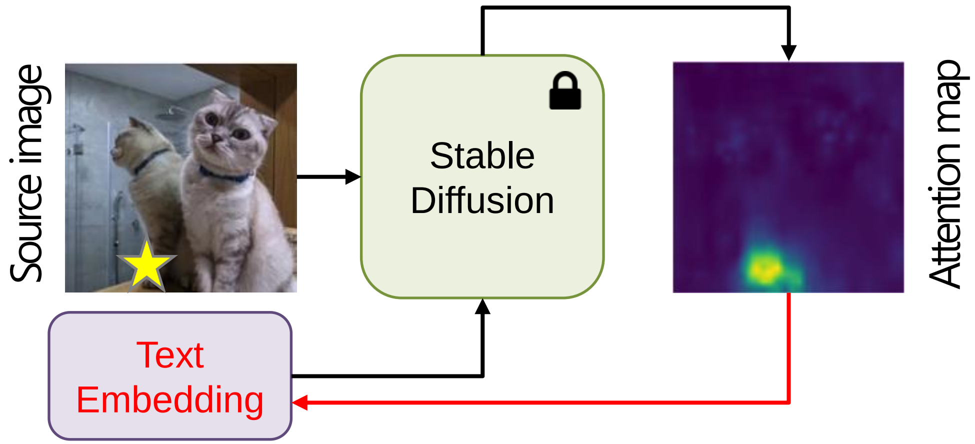
Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Method

사전 학습된 Stable Diffusion의 **어텐션 맵**을 통해 특정 위치에 대해 **최적화된 prompt embedding**을 활용하여 correspondence task 수행

1. 쿼리 위치에 대한 최적의 임베딩을 찾음



<Step 1>

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Method 1

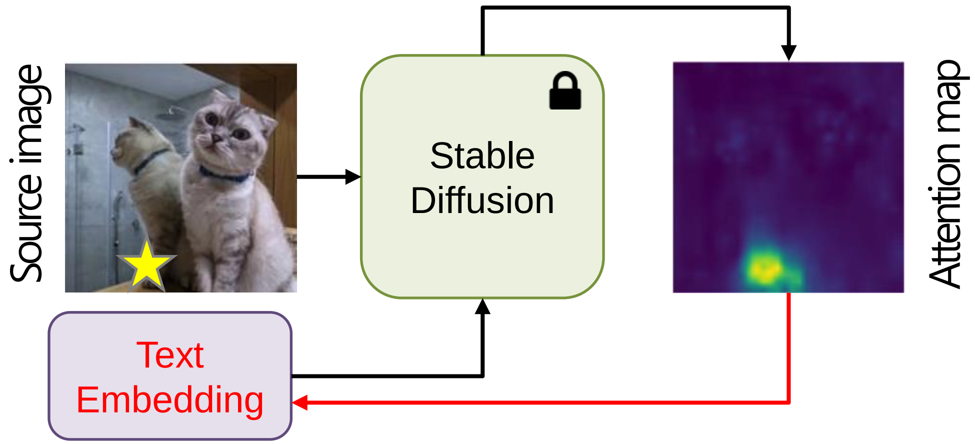
Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Method

사전 학습된 Stable Diffusion의 **어텐션 맵**을 통해 특정 위치에 대해 **최적화된 prompt embedding**을 활용하여 correspondence task 수행

1. 쿼리 위치에 대한 최적의 임베딩을 찾음　　Textual-Inversion 방식과 유사하게 진행

   ✓ 최적화된 임베딩은 위치에 대한 semantic information을 담고 있음



<Step 1>

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.
Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-or, D. (2022, September). An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In The Eleventh International Conference on Learning Representations.

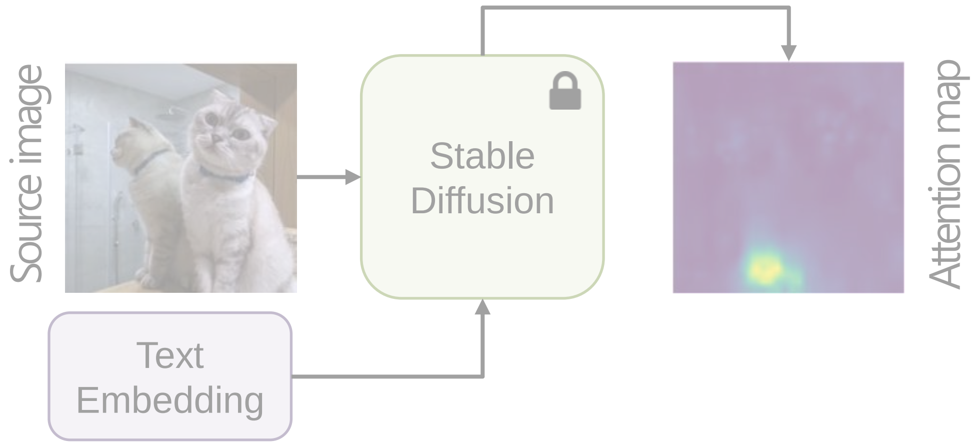
Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

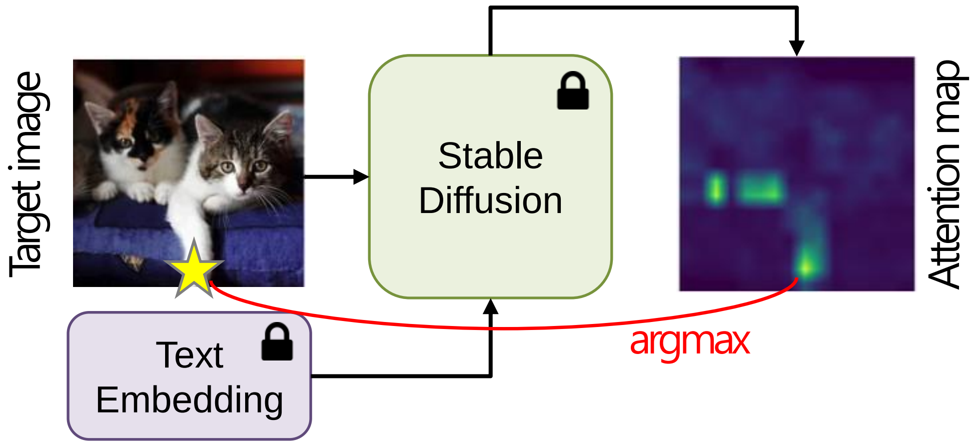
❖ Method

사전 학습된 Stable Diffusion의 **어텐션 맵**을 통해 특정 위치에 대해 **최적화된 prompt embedding**을 활용하여 correspondence task 수행

1. 쿼리 위치에 대한 최적의 임베딩을 찾음
2. Target image에 해당 임베딩을 적용하여 argmax인 부분을 추출 → semantic correspondence



<Step 1>



<Step 2>

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Method 1

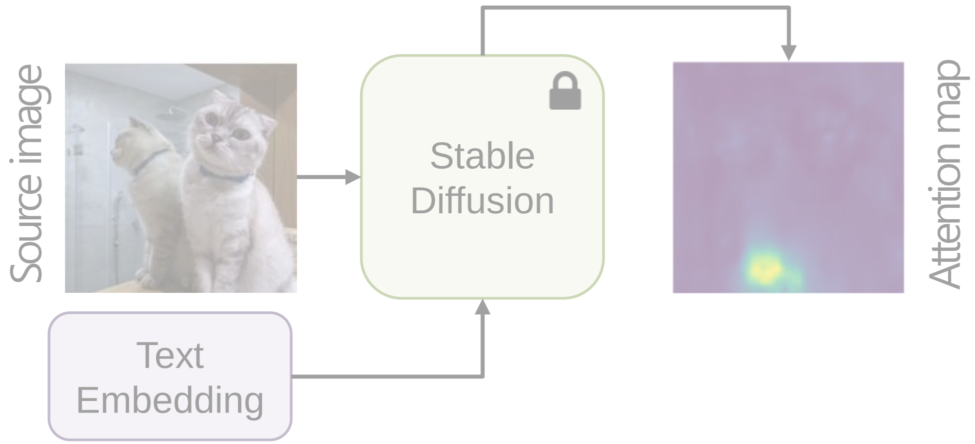
Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Method

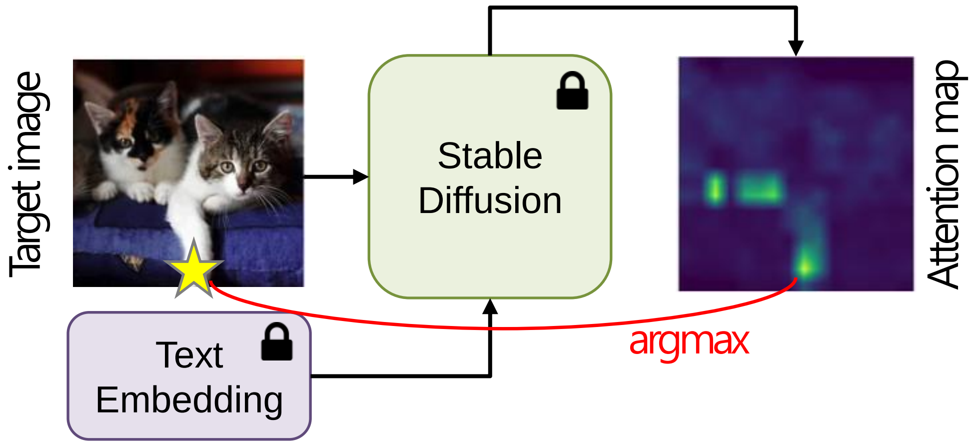
사전 학습된 Stable Diffusion의 **어텐션 맵**을 통해 특정 위치에 대해 **최적화된 prompt embedding**을 활용하여 correspondence task 수행

1. 쿼리 위치에 대한 최적의 임베딩을 찾음
2. Target image에 해당 임베딩을 적용하여 argmax인 부분을 추출 → semantic correspondence

Attention maps을 사용하지만,
실제 단어에 의존하지 않음



<Step 1>



<Step 2>

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Method 1

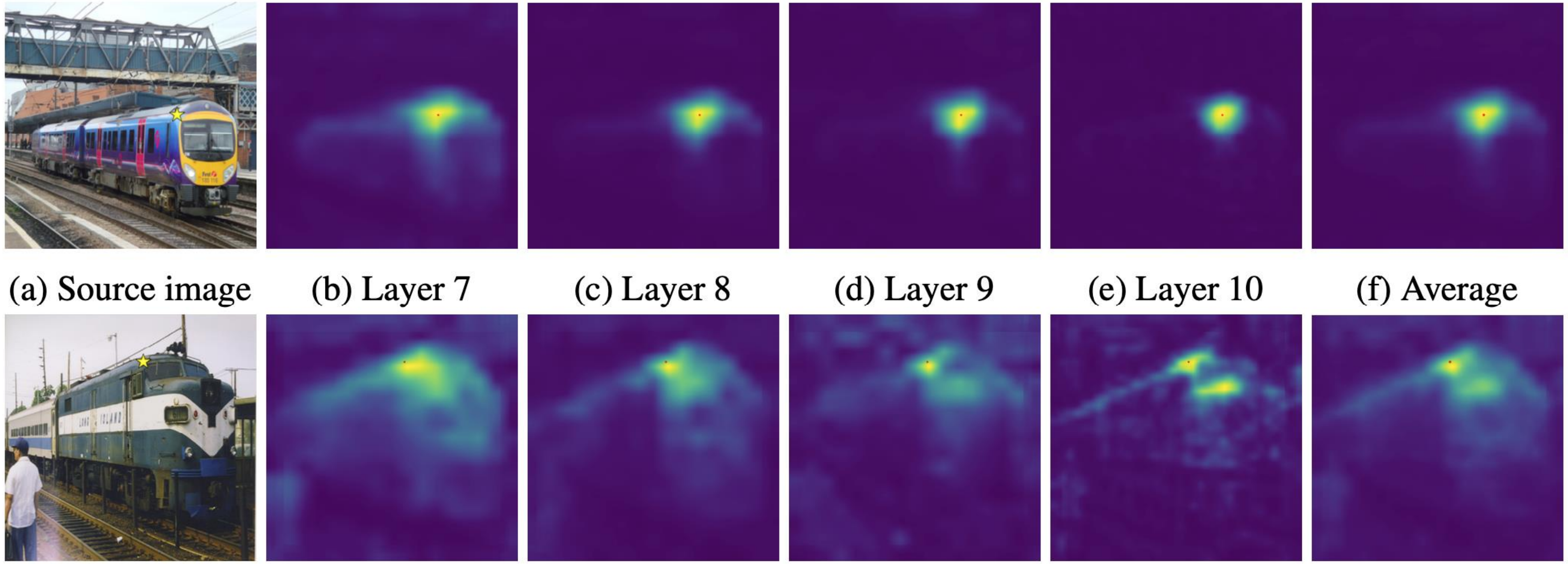
Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Method

사전 학습된 Stable Diffusion의 **어텐션 맵**을 통해 특정 위치에 대해 **최적화된 prompt embedding**을 활용하여 correspondence task 수행

- Attention response of different layers

✓ 각 layer의 다른 특성들을 활용하기 위해 **(f) Average** 사용



(a) Source image (b) Layer 7 (c) Layer 8 (d) Layer 9 (e) Layer 10 (f) Average

(g) Target image (h) Layer 7 (i) Layer 8 (j) Layer 9 (k) Layer 10 (l) Average

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Experiments

1) Semantic Correspondence - Quantitative

✓ **Metric**: Percentage of Correct Keypoints(PCK)

		CUB-200		PF-Willow		SPair-71k	
		$PCK_{@0.05}$	$PCK_{@0.1}$	$PCK_{@0.05}$	$PCK_{@0.1}$	$PCK_{@0.05}$	$PCK_{@0.1}$
Strong supervision	PWarpC-NC-Net*$_{res101}$	-	-	48.0	76.2	21.5	37.1
	CHM	-	-	52.7	79.4	27.2	46.3
	VAT	-	-	52.8	81.6	35.0	55.5
	CATs++	-	-	56.7	81.2	–	59.8
Weak supervision	PMD	-	-	40.3	74.7	–	26.5
	PSCNet-SE	-	-	42.6	75.1	–	27.0
	VGG+MLS	18.3	25.8	41.2	63.2	–	27.4
	DINO+MLS	52.0	67.0	45.0	66.5	–	31.1
	PWarpC-NC-Net$_{res101}$	–	–	45.0	75.9	18.2	35.3
	ASIC	57.9	75.9	**53.0**	76.3	–	36.9
Unsupervised	DINO+NN	52.8	68.3	40.1	60.1	–	33.3
	Our method	**61.6**	**77.5**	**53.0**	**84.3**	**28.9**	**45.4**

threshold

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

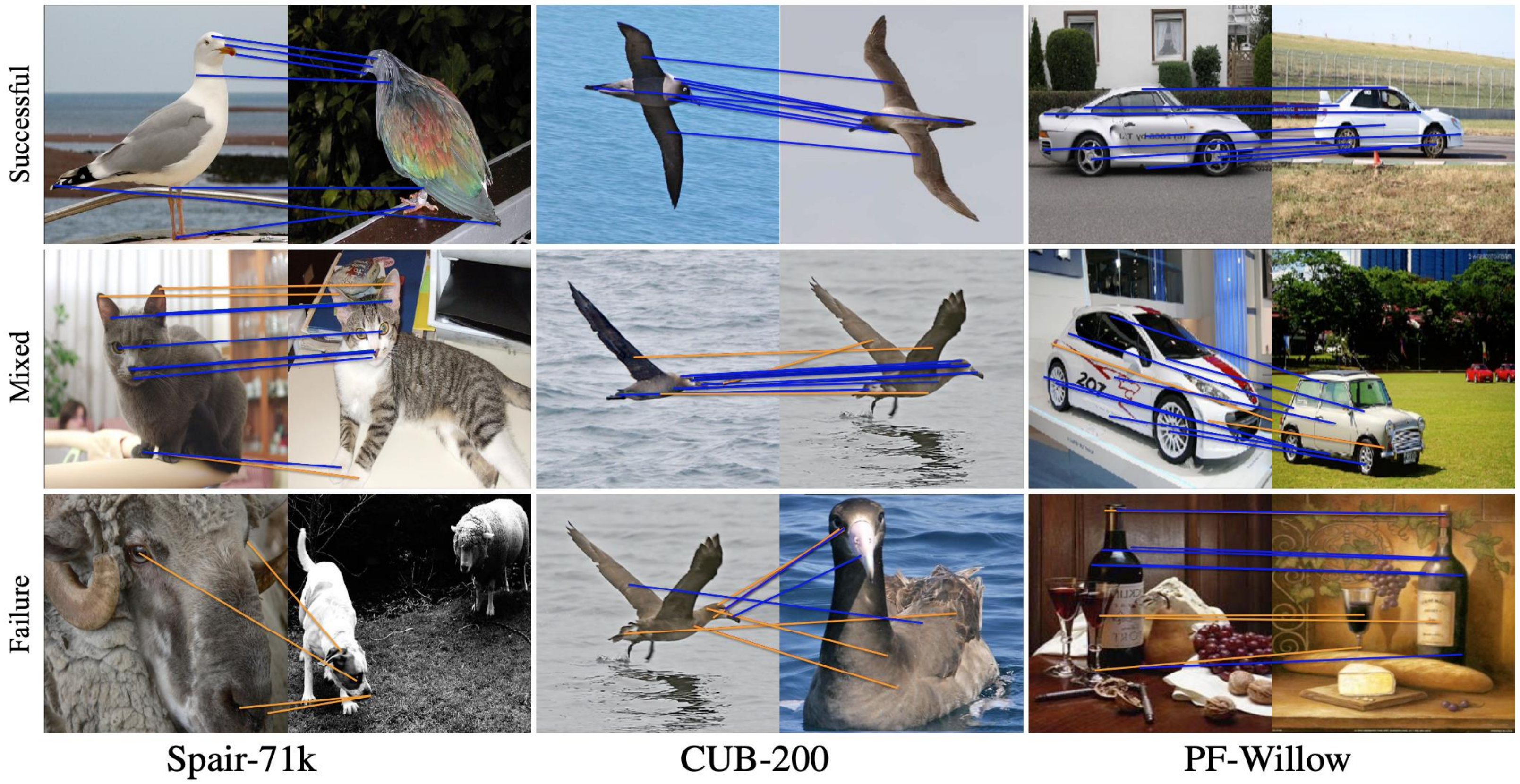
Method 1

Unsupervised Semantic Correspondence Using Stable Diffusion

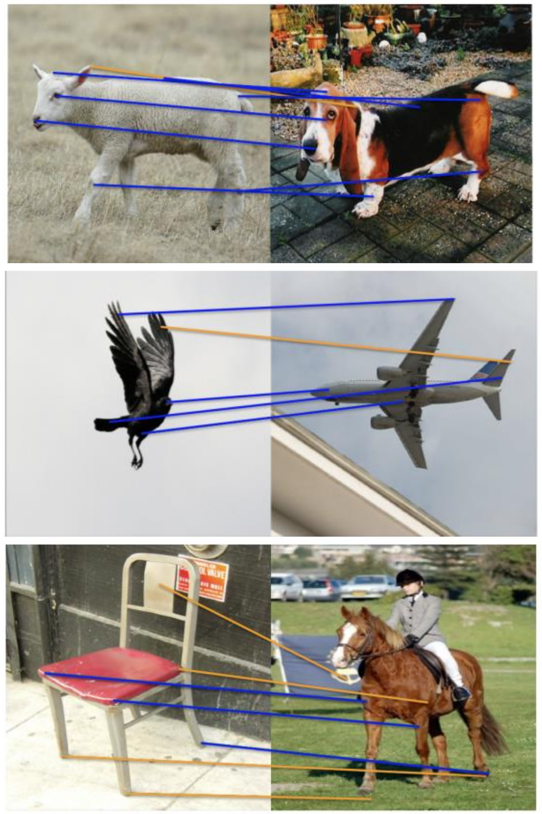
❖ Experiments

2) Semantic Correspondence - Qualitative

Blue : Correct
Orange : Wrong



<Between **same** classes>



<Between **different** classes>

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Methods

Text-to-image diffusion models

Unsupervised
Semantic Correspondence
Using Stable Diffusion

Image-to-image diffusion models

Emergent Correspondence
from Image Diffusion

Image-to-image diffusion models

Diffusion Hyperfeatures:
Searching Through
Time and Space
for Semantic Correspondence

Method 2

Emergent Correspondence from Image Diffusion

❖ Emergent Correspondence from Image Diffusion(NeurIPS, 2023)

- **Motivation**: Diffusion Models이 Image Editing에 좋은 성능을 보이는 것은 이미지 간 correspondence을 추론하기 때문이 아닐까?

Cat → Dog

Without changing
its pose or context

Image-to-Image Translation
(pix2pix-zero)

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., & Zhu, J. Y. (2023, July). Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings* (pp. 1-11).

Method 2

Emergent Correspondence from Image Diffusion

❖ Emergent Correspondence from Image Diffusion(NeurIPS, 2023)

- **Motivation**: Diffusion Models이 Image Editing에 좋은 성능을 보이는 것은 이미지 간 correspondence을 추론하기 때문이 아닐까?

Cat → Dog

Without changing its pose or context

→ Implicit correspondence?

Image-to-Image Translation
(pix2pix-zero)

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., & Zhu, J. Y. (2023, July). Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings* (pp. 1-11).

Method 2

Emergent Correspondence from Image Diffusion

❖ Emergent Correspondence from Image Diffusion(NeurIPS, 2023)

- **Motivation**: Diffusion Models이 Image Editing에 좋은 성능을 보이는 것은 이미지 간 correspondence을 추론하기 때문이 아닐까?
- 사전 학습된 Diffusion Models을 사용해 correspondence을 추출할 수 있는 간단한 방법론 제안

Cat → Dog

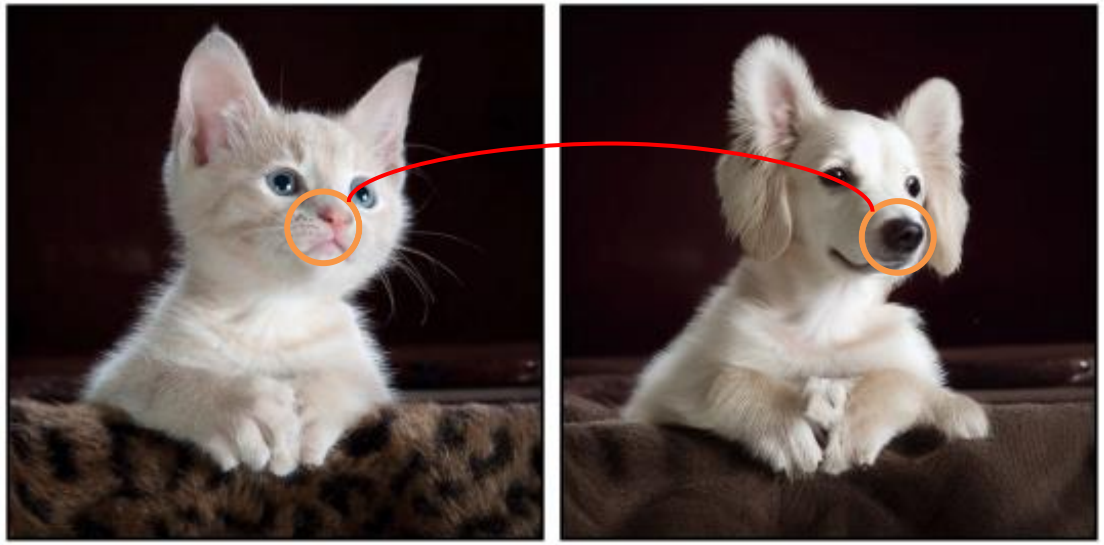


Image-to-Image Translation
(pix2pix-zero)

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., & Zhu, J. Y. (2023, July). Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings* (pp. 1-11).

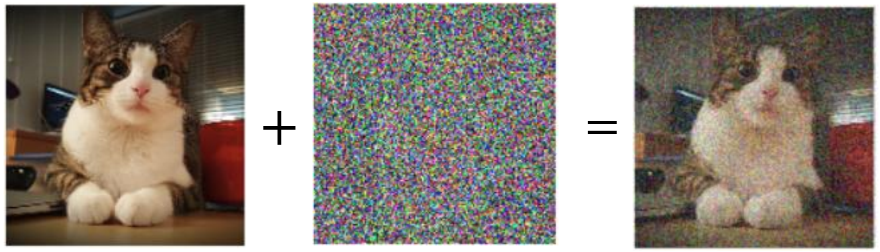
Method 2

Emergent Correspondence from Image Diffusion

❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

1. Input image x_0에 noise 더해 x_t 생성



x_0 $\qquad$ ϵ $\qquad$ x_t

$$x_t = \sqrt{\alpha_t}x_0 + \left(\sqrt{1-\alpha_t}\right)\epsilon,\ \epsilon \sim N(0, I)$$

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

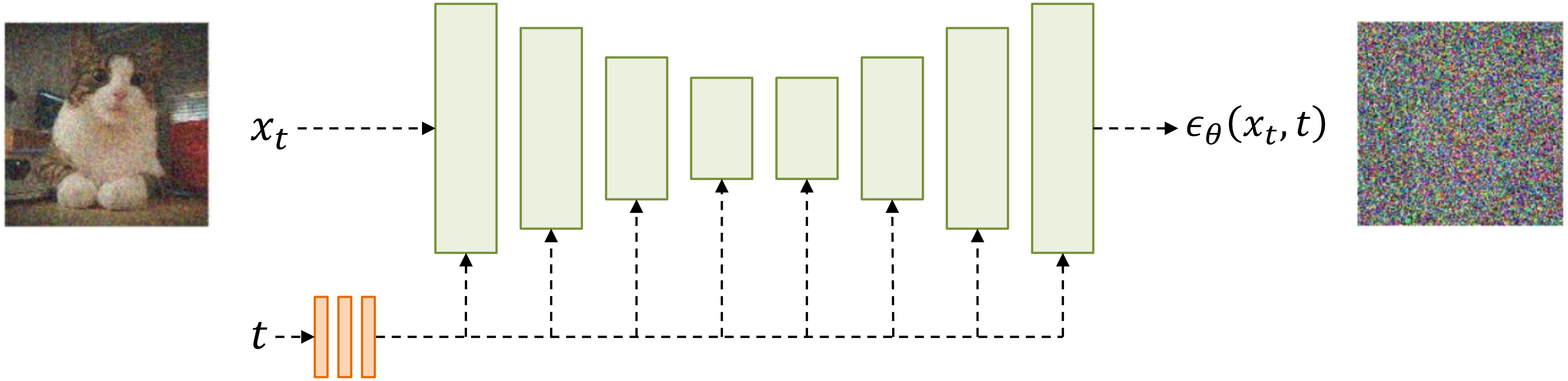
Method 2

Emergent Correspondence from Image Diffusion

❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

2. Denoising network $\epsilon_\theta(x_t, t)$에 noisy image x_t와 timestep t를 입력



Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Method 2

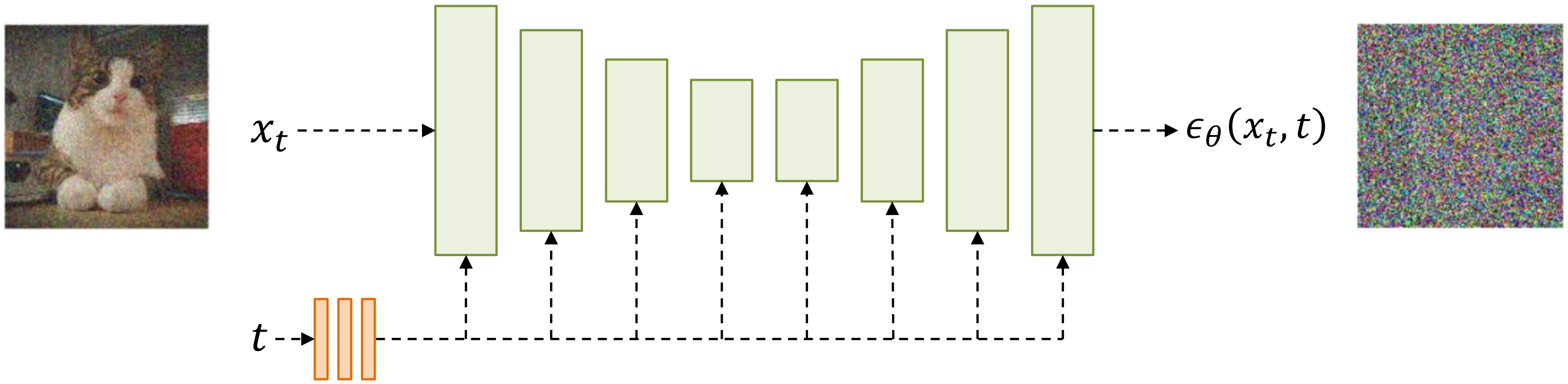
Emergent Correspondence from Image Diffusion

❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

2. Denoising network $\epsilon_\theta(x_t, t)$에 noisy image x_t와 timestep t를 입력

   ✓ Denoising network로 사전 학습된 U-Net의 입력으로 사용하기 위해 원본 이미지가 아닌 노이지한 이미지 생성 후 입력



Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

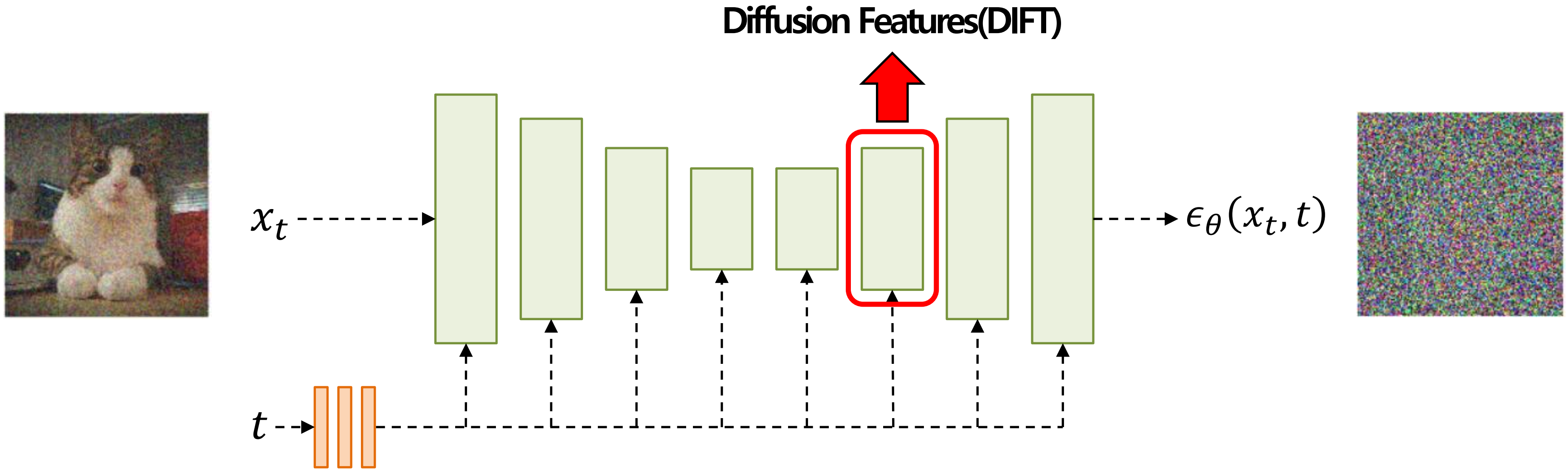
Method 2

Emergent Correspondence from Image Diffusion

❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

3. Denoising 과정 중 U-Net의 특정 upsampling block i에서 **intermediate feature maps**을 추출



Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

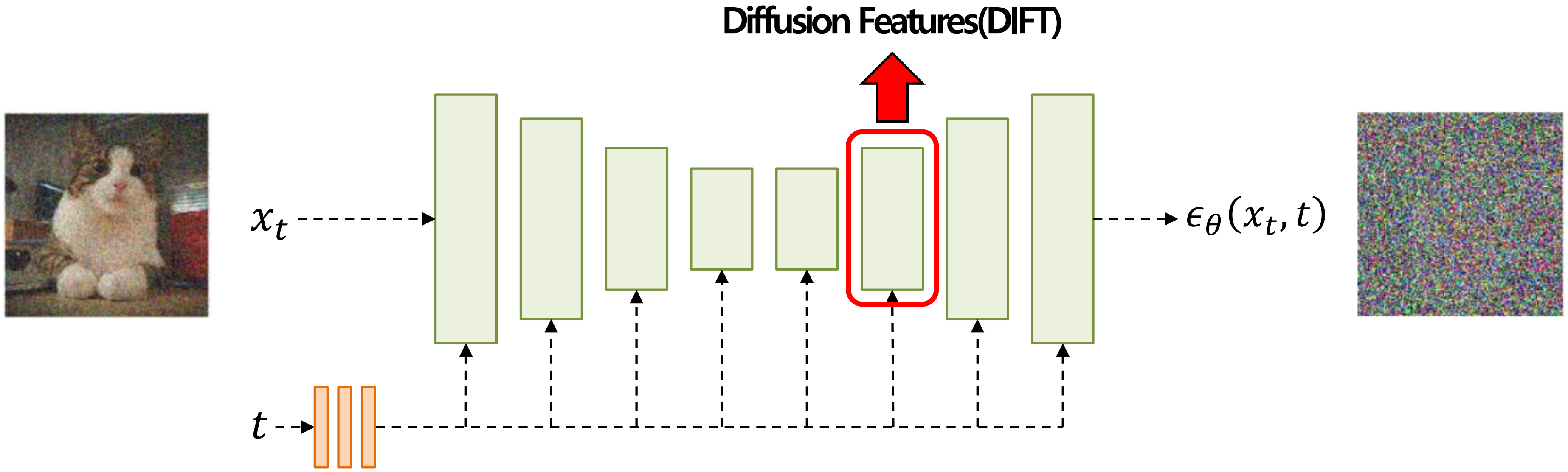
Method 2

Emergent Correspondence from Image Diffusion

❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

3. Denoising 과정 중 특정 block i에서 **intermediate feature maps**을 추출 → Diffusion Features(DIFT)



Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

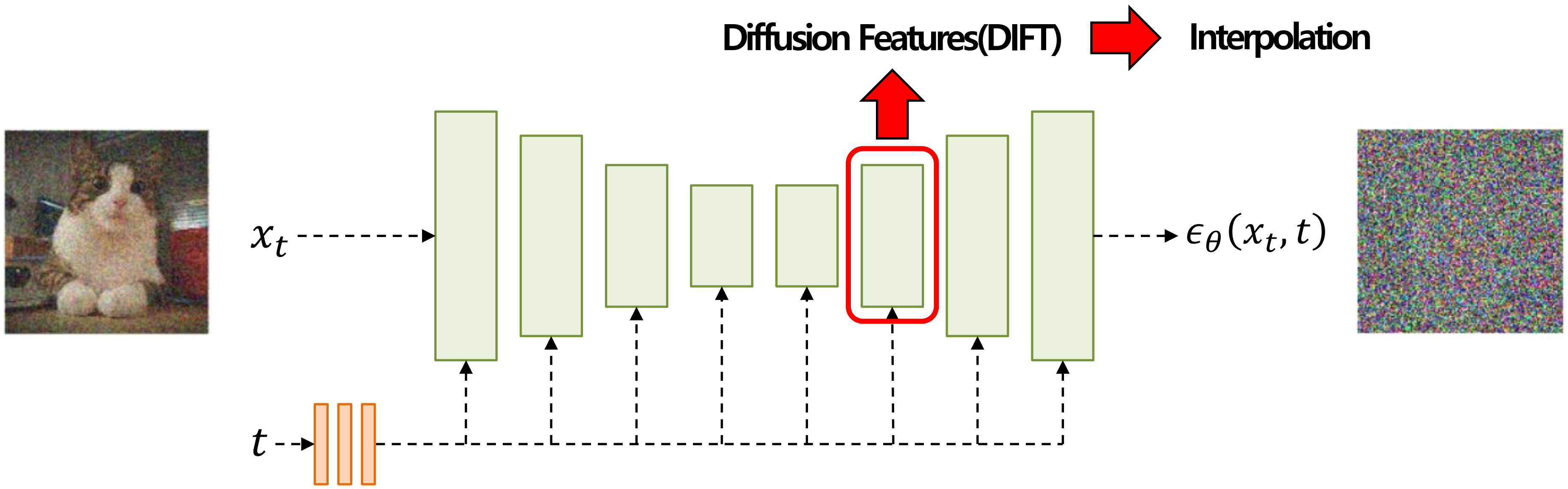
Method 2

Emergent Correspondence from Image Diffusion

❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

4. 각 포인트 feature vectors을 얻기 위해 Interpolation 진행



Method 2

Emergent Correspondence from Image Diffusion

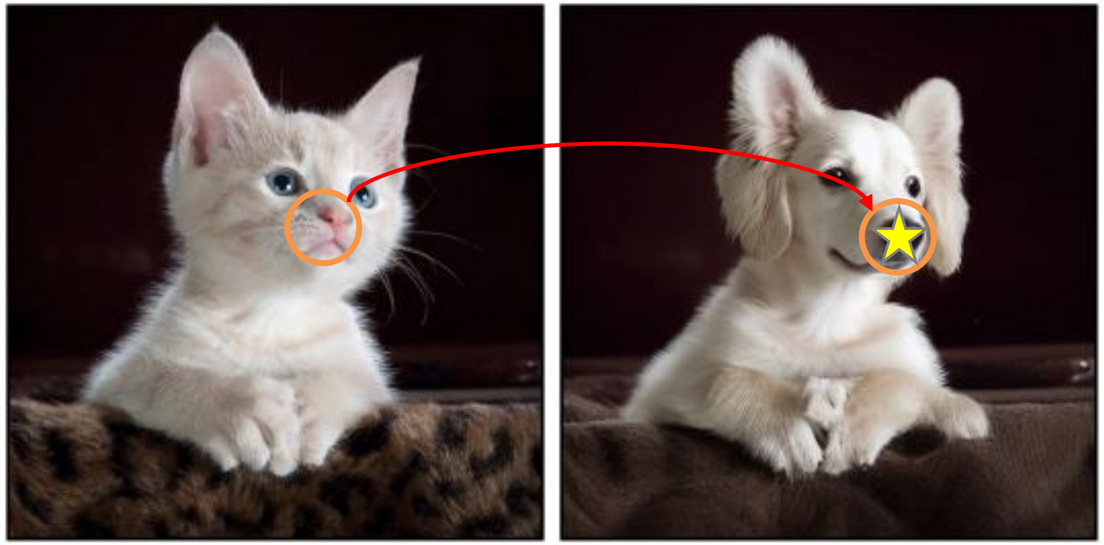
❖ Method

사전 학습된 Diffusion Models 통해 추출한 Input image의 **image features**을 활용하여 correspondence task 수행

5. 두 이미지의 Feature matching(cosine similarity)을 통해 correspondences을 구함

$$p_2 = \underset{p}{\operatorname{argmin}}\, d(F_1(p_1), F_2(p))$$

Cat → Dog



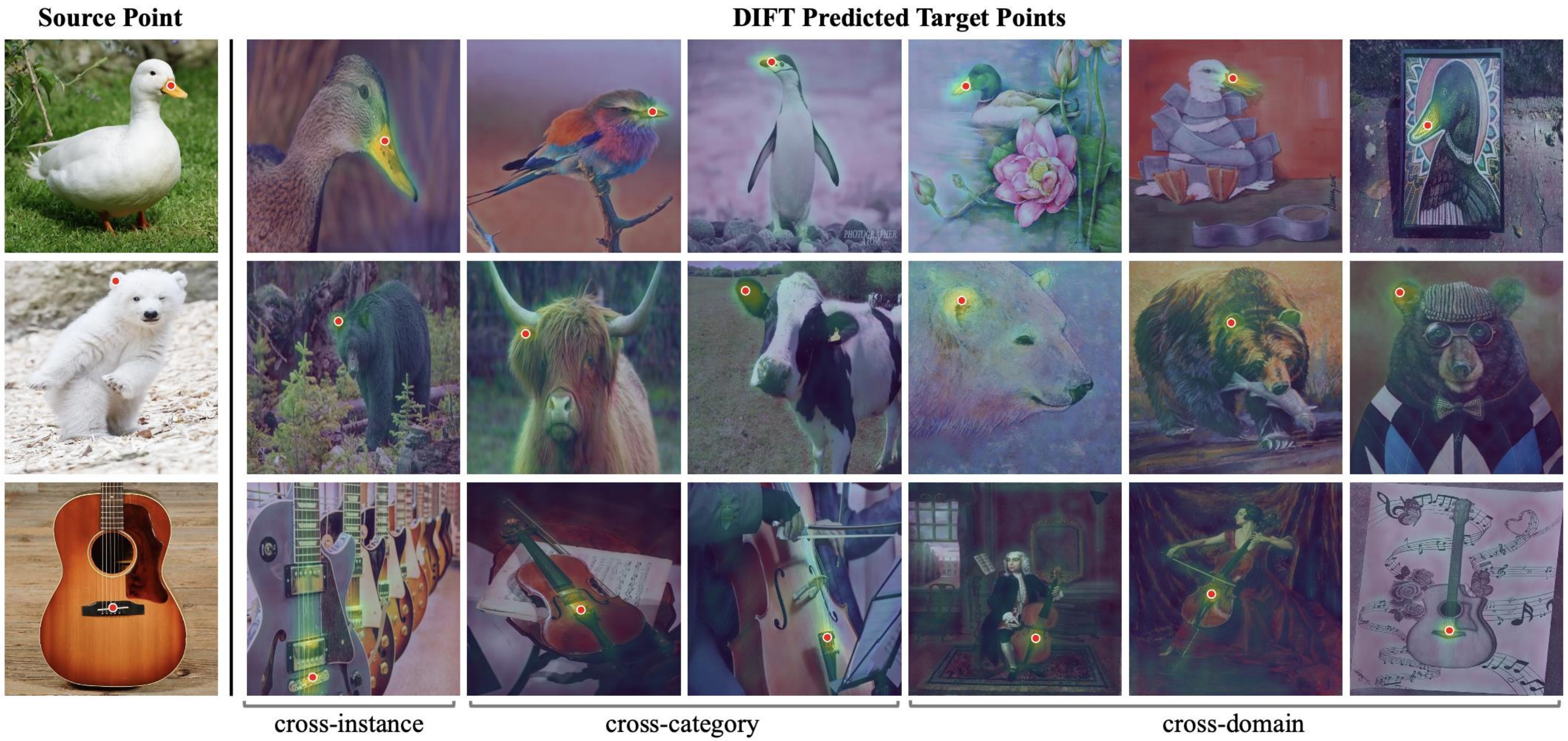
Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Method 2

Emergent Correspondence from Image Diffusion

❖ Experiments

1) Semantic Correspondence



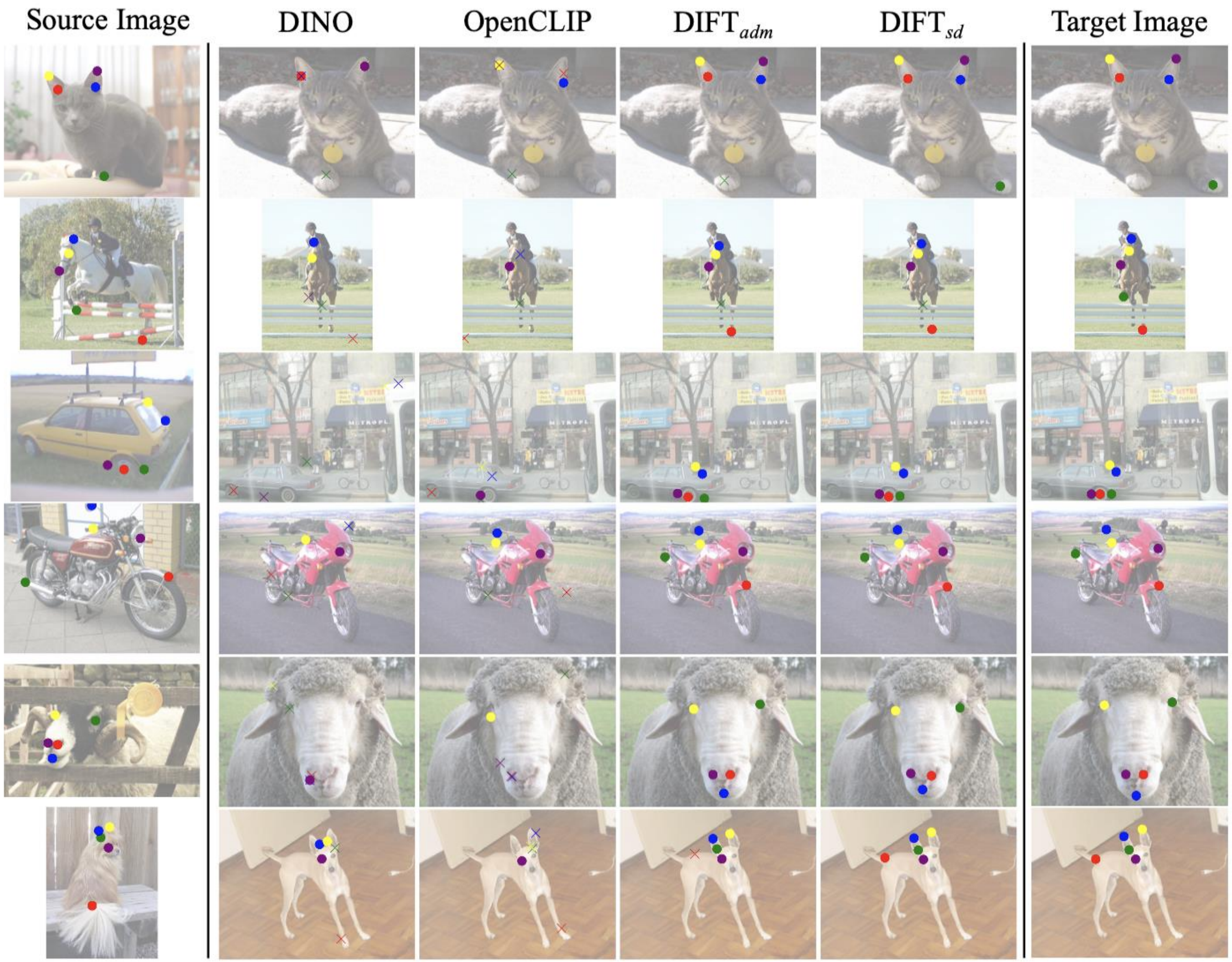
Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Method 2

Emergent Correspondence from Image Diffusion

❖ Experiments

1) Semantic Correspondence
   - ✓ Cluttered scenes
   - ✓ Viewpoint changes
   - ✓ Occlusions



Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Method 2

Emergent Correspondence from Image Diffusion

❖ Experiments

1) Semantic Correspondence

✓ Across various categories($DIFT_{sd}$)

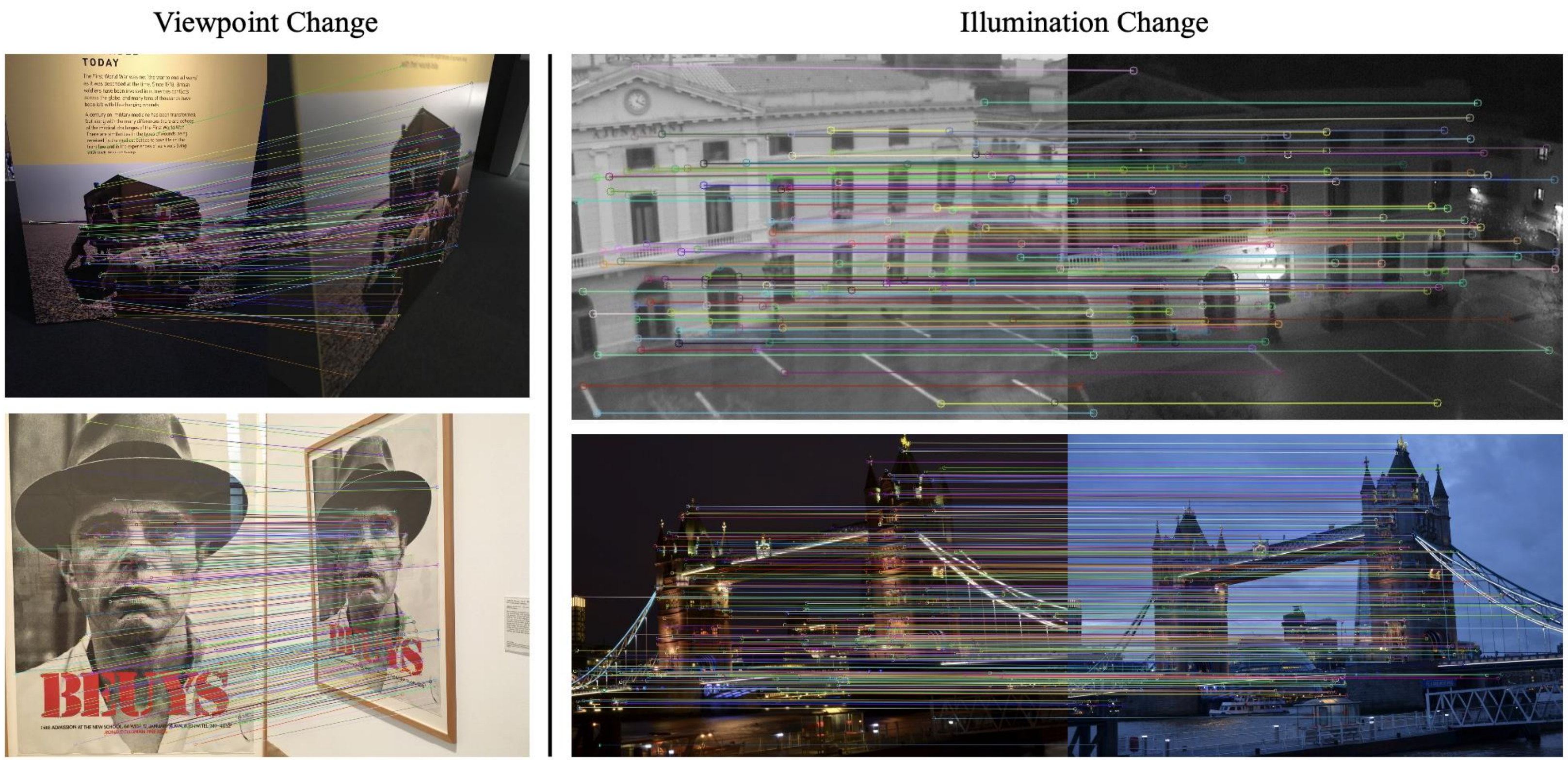
Method 2

Emergent Correspondence from Image Diffusion

❖ Experiments

2) Geometric Correspondence

- ✓ Viewpoint Change
- ✓ Illumination Change



Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

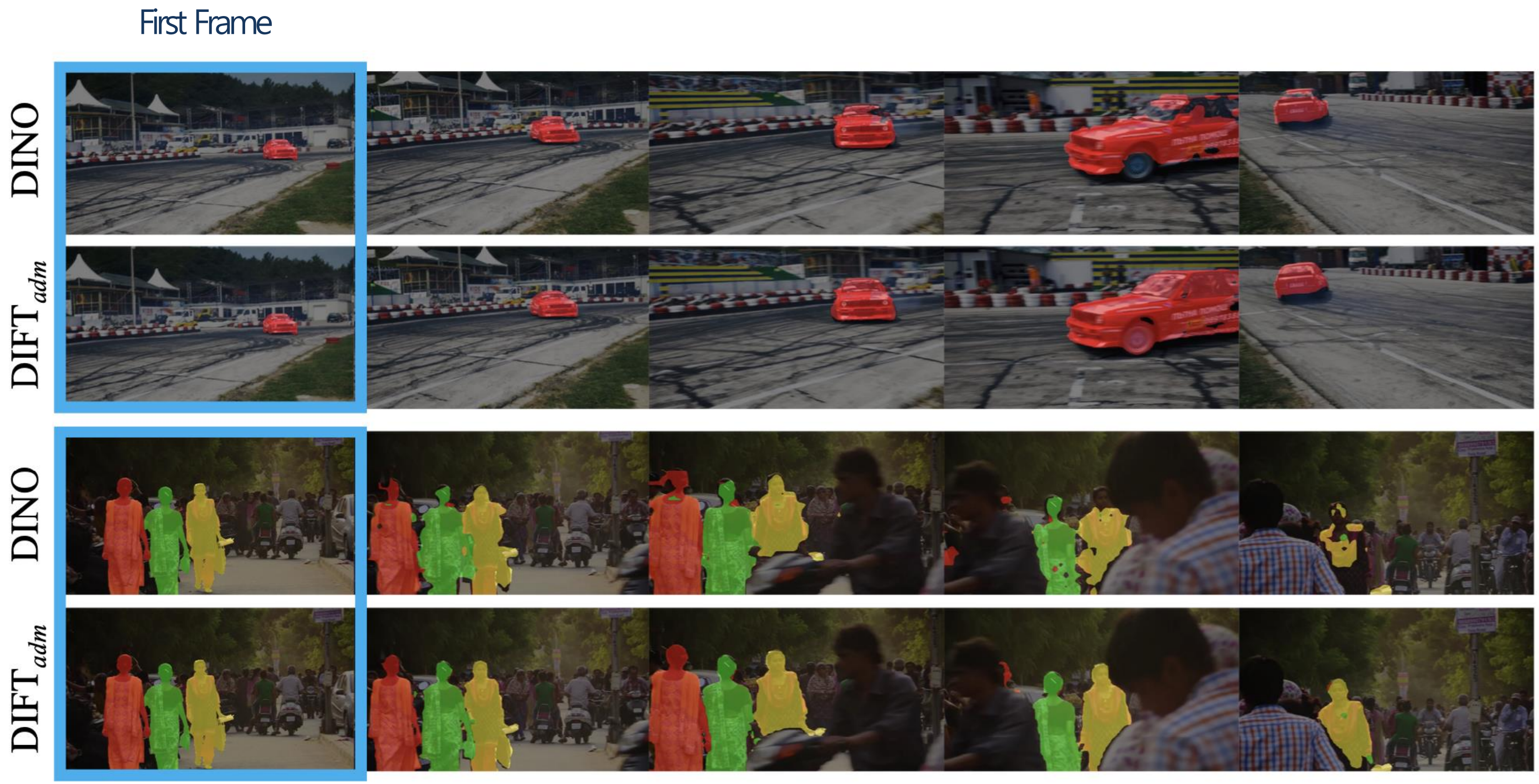
Method 2

Emergent Correspondence from Image Diffusion

❖ **Experiments**

3) Temporal Correspondence

✓ Video label propagation



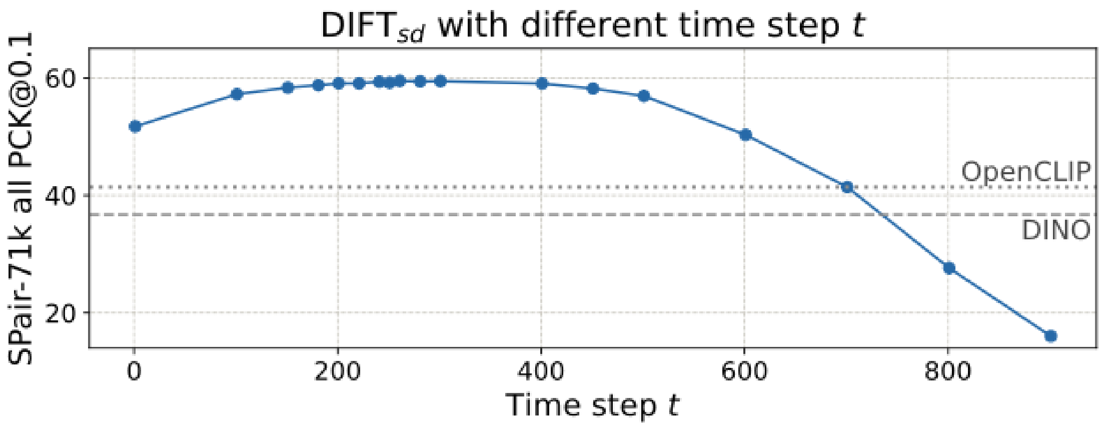
Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Method 2

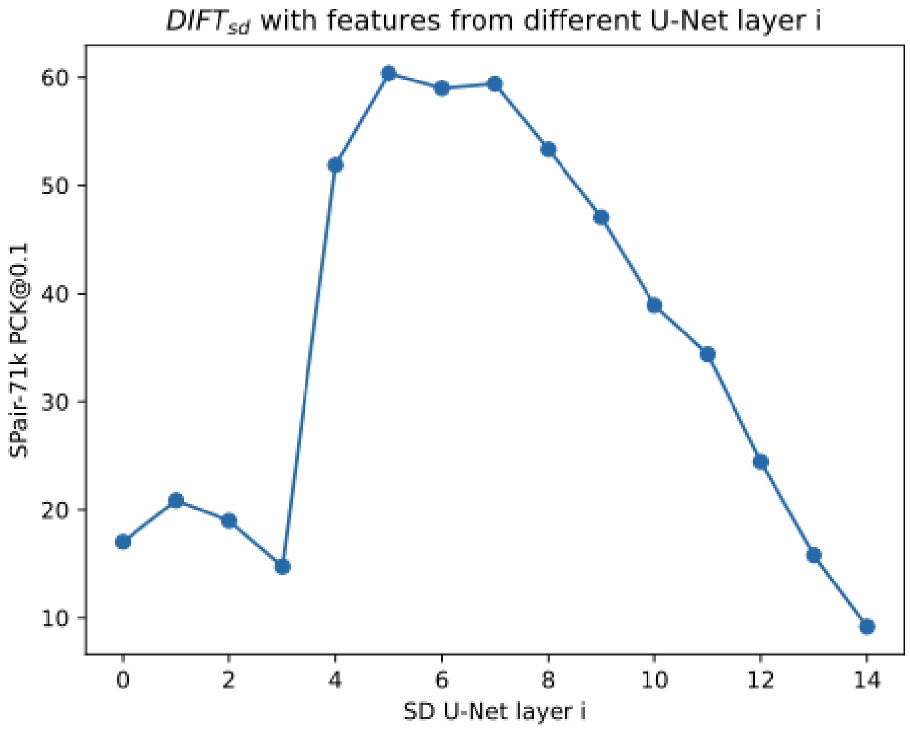
Emergent Correspondence from Image Diffusion

❖ Experiments

4) Ablation – Time step & U-Net layer



Time step



SD U-Net layer

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Methods

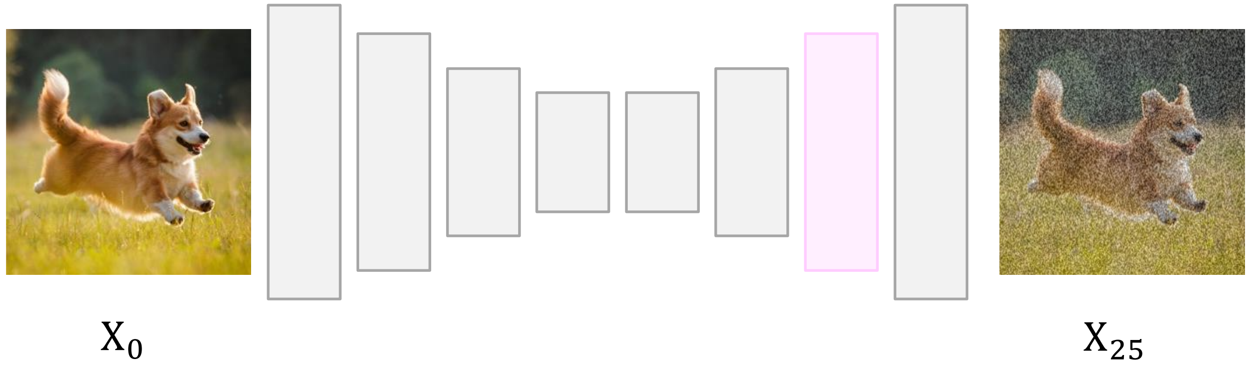
Text-to-image diffusion models	Image-to-image diffusion models	Image-to-image diffusion models
Unsupervised Semantic Correspondence Using Stable Diffusion	Emergent Correspondence from Image Diffusion	Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ **Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence(NeurIPS, 2023)**

- **Motivation**: 특정 시점과 레이어를 선택하는 것이 아니라 전체를 다 사용하면 더 좋지 않을까?



특정 시점 & 특정 레이어

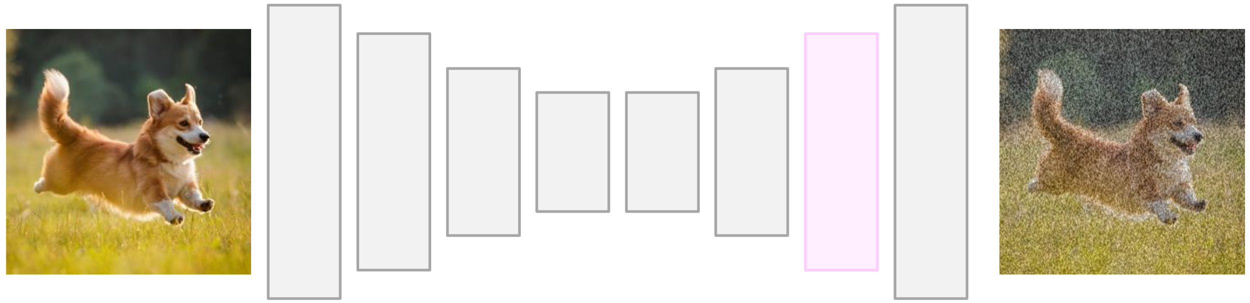
추가적인 과정이 필요하며,
다른 features에 있는 정보들을 활용하지 못함

Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ **Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence(NeurIPS, 2023)**

- **Motivation**: 특정 시점과 레이어를 선택하는 것이 아니라 전체를 다 사용하면 더 좋지 않을까?



X_0 X_{25}

특정 시점 & 특정 레이어

추가적인 과정이 필요하며,
다른 features에 있는 정보들을 활용하지 못함

→ 특정 시점, 특정 레이어의 Features를 사용하지 말고,
전체를 활용할 수 있는 구조 제안

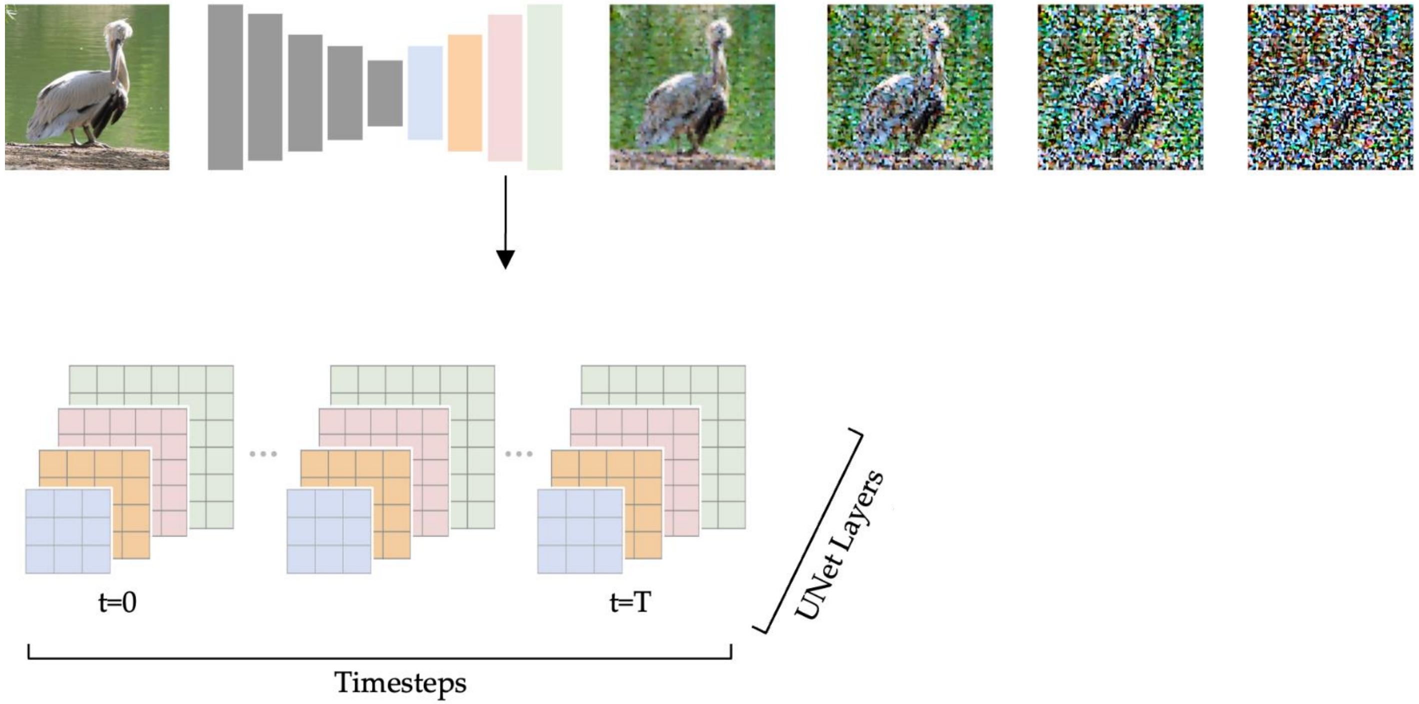
Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Method

Diffusion Process의 **전체 intermediate feature maps 통합**하는 프레임워크를 제안

1. Diffusion process 중 발생하는 모든 feature maps 저장



Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

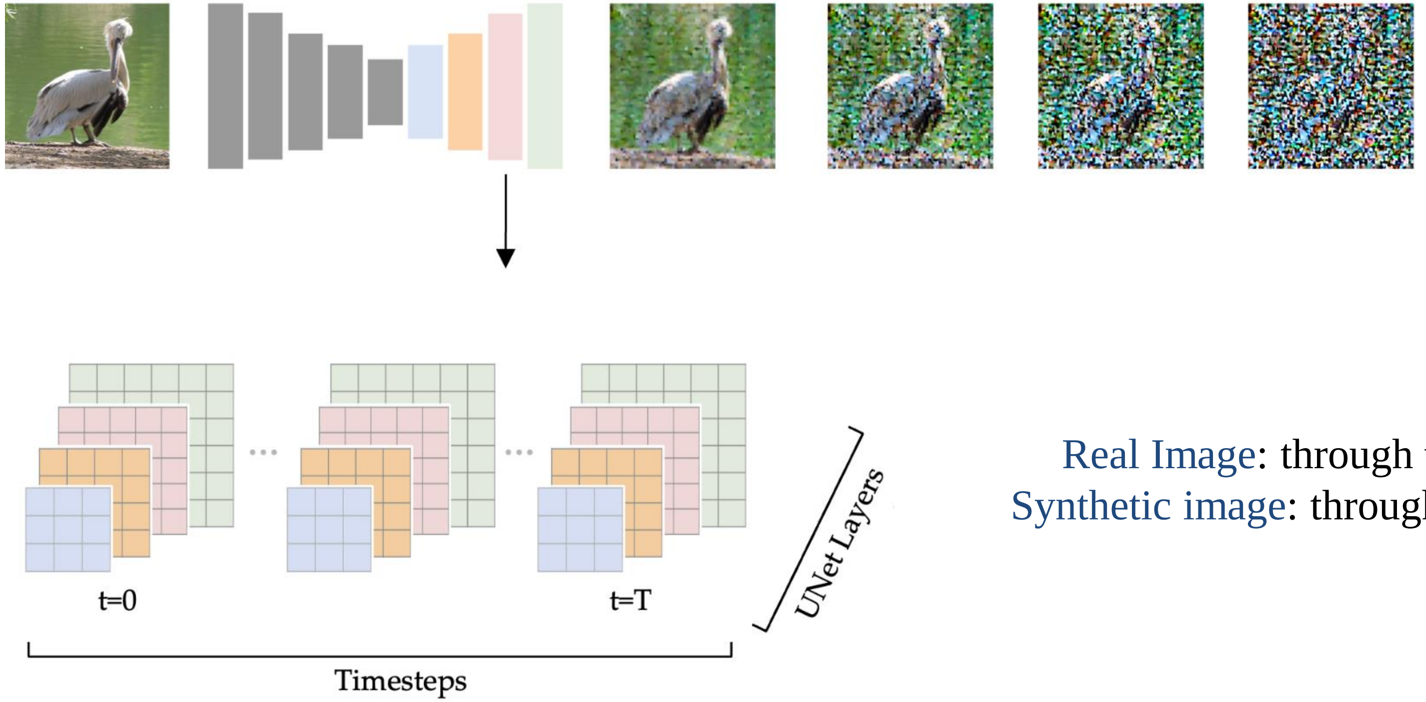
Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Method

Diffusion Process의 **전체 intermediate feature maps 통합**하는 프레임워크를 제안

1. Diffusion process 중 발생하는 모든 feature maps 저장



Real Image: through the <u>inversion</u> process
Synthetic image: through the <u>generation</u> process

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

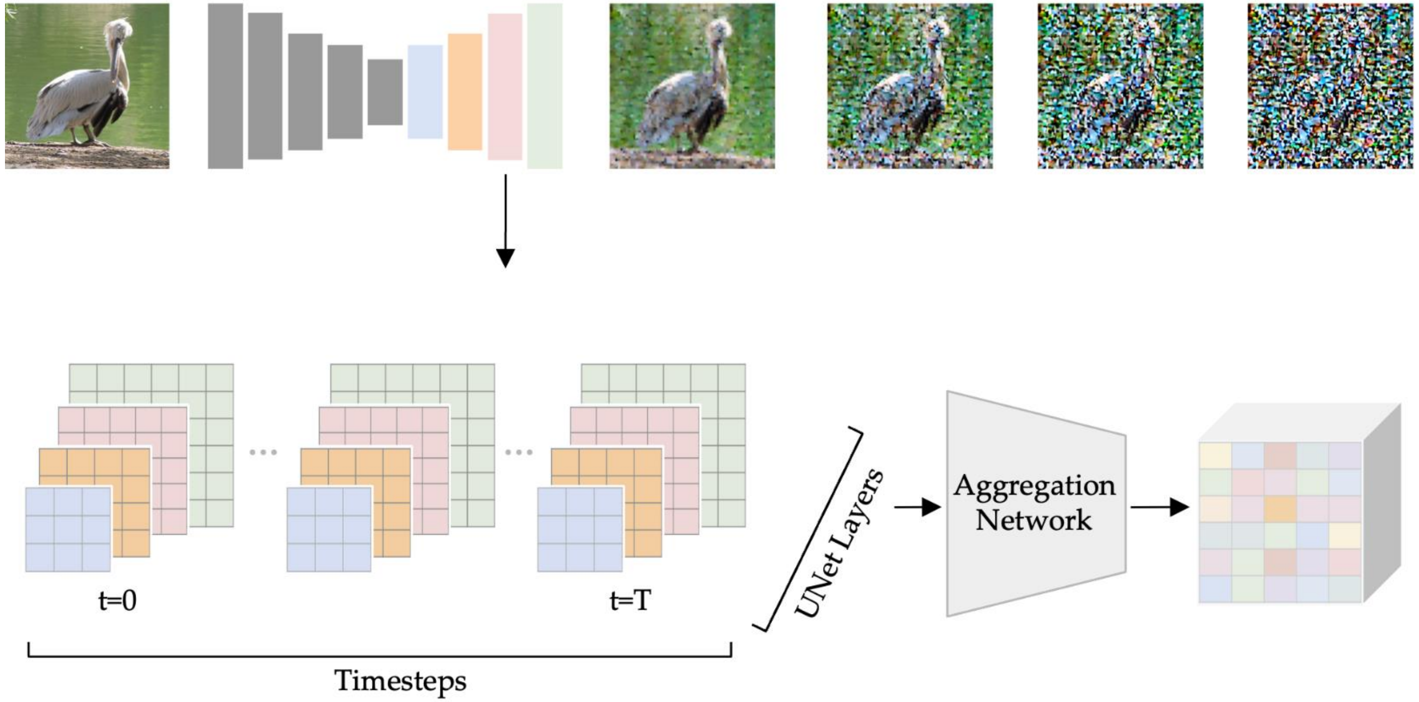
Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Method

Diffusion Process의 **전체 intermediate feature maps 통합**하는 프레임워크를 제안

2. Aggregation Network를 통해 모든 feature maps를 통합하여 a single feature map 생성



Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

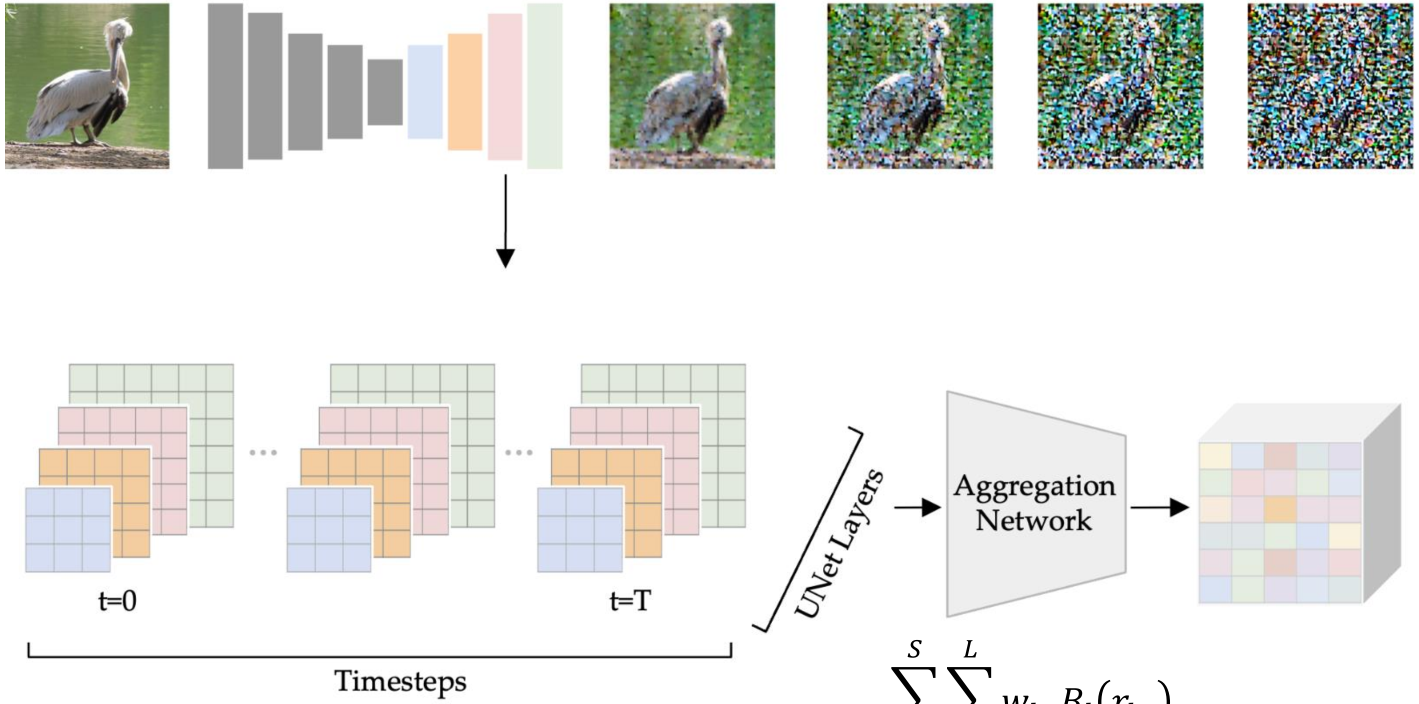
Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Method

Diffusion Process의 **전체 intermediate feature maps 통합**하는 프레임워크를 제안

2. Aggregation Network를 통해 모든 feature maps를 통합하여 a single feature map 생성



$$\sum_{s=0}^{S}\sum_{l=1}^{L} w_{l,s} B_l(r_{l,s})$$

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

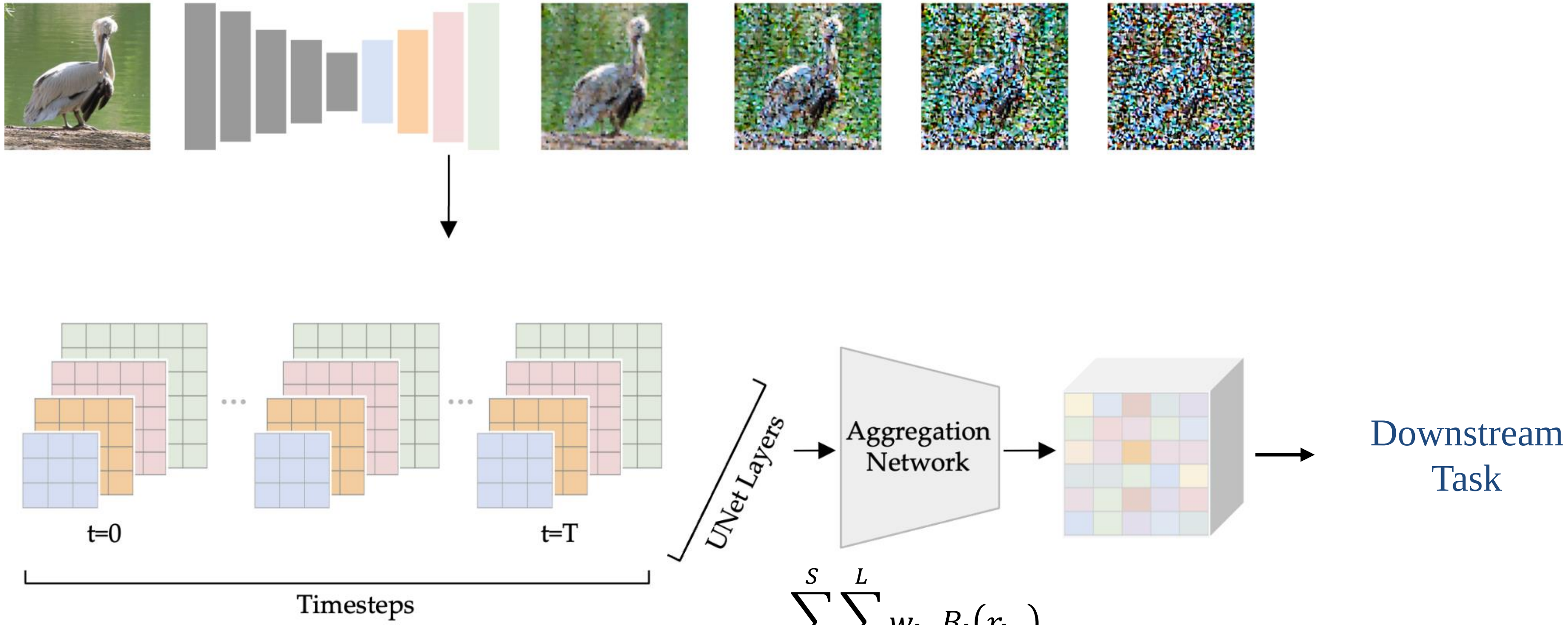
Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Method

Diffusion Process의 **전체 intermediate feature maps 통합**하는 프레임워크를 제안

3. Downstream task에 맞게 학습 (Semantic Correspondence: by performing a nearest neighbor searching)



$$\sum_{s=0}^{S}\sum_{l=1}^{L} w_{l,s} B_l(r_{l,s})$$

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

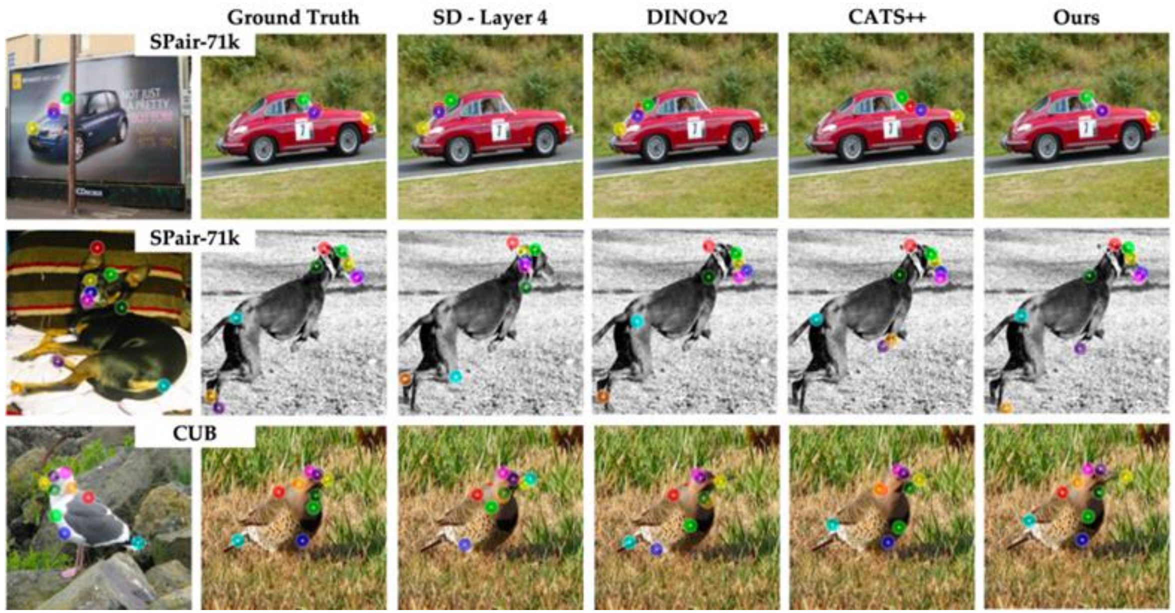
Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Experiments

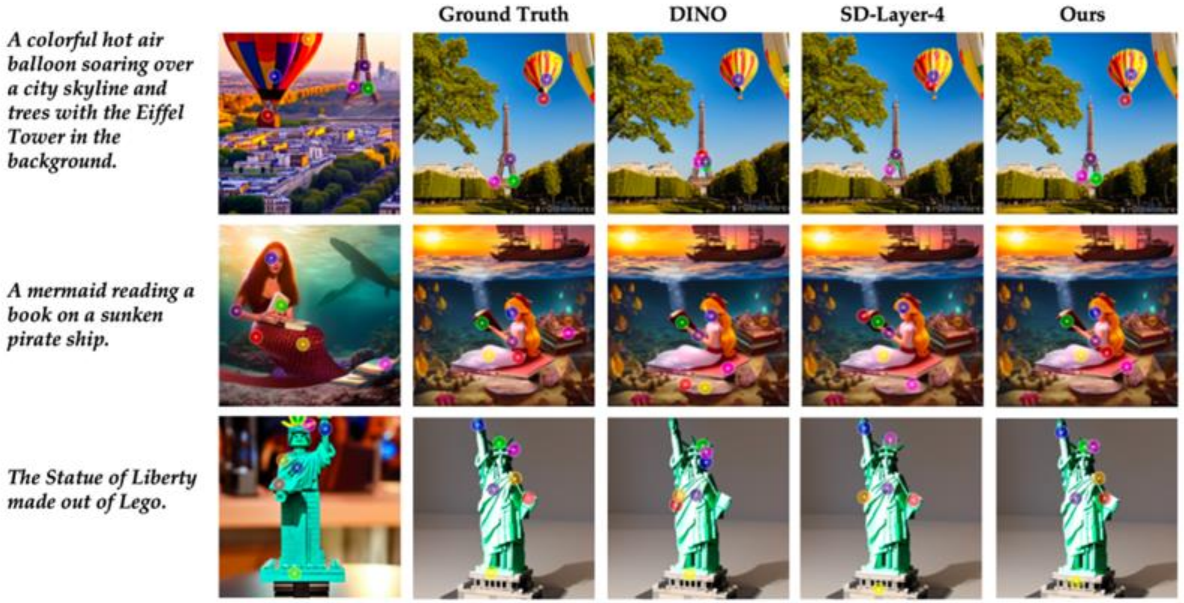
1) Semantic Correspondence - Qualitative

Single layer Self-supervised Supervised



Real images

Self-supervised Single layer



Synthetic images

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Experiments

1) Semantic Correspondence - Quantitative

		# Layers L	# Timesteps S	SPair-71k ↑ PCK@0.1$_{img}$	SPair-71k ↑ PCK@0.1$_{bbox}$	CUB ↑ PCK@0.1$_{img}$	CUB ↑ PCK@0.1$_{bbox}$
Self-supervised	DINO [2]	1	-	51.68	41.04	72.72	55.90
	DINOv2 [32]	1	-	68.33	56.98	*89.96**	*76.83**
Supervised	DHPF [29]	34	-	55.28	42.63	77.30	61.42
	CATS++ [6]	30	-	70.26	57.06	75.92	59.49
	SD-Layer-4	1	1	58.80	46.58	78.43	61.22
	SD-Concat-All	12	1	52.12	41.83	70.22	54.05
	Ours	12	11	**72.56**	**64.61**	**82.29**	**69.42**
	Ours-One-Step	12	1	63.74	54.69	76.59	62.11
	SD-Layer-Pruned	1	1	57.69	48.16	80.67	67.21
	Ours-Pruned	1	1	64.02	53.74	79.10	63.95
	Ours-SDv2-1	12	11	70.74	64.85	80.39	68.04

Table 1: We evaluate our semantic keypoint matching performance on real images from SPair-71k and CUB. For our CUB evaluation, we transfer the model tuned on SPair-71k. We compare against Stable Diffusion baselines that extract features from a single layer (SD-Layer-4) or concatenation of all layers (SD-Concat-All). We ablate pruning to the single feature map with the highest mixing weight selected by our method, either as the raw feature map (SD-Layer-Pruned) or after the bottleneck layer (Ours-Pruned). We ablate tuning our method with only one timestep (One-Step) or features from another Stable Diffusion variant (SDv2-1). **Note that DINOv2 was trained on samples from CUB [32].*

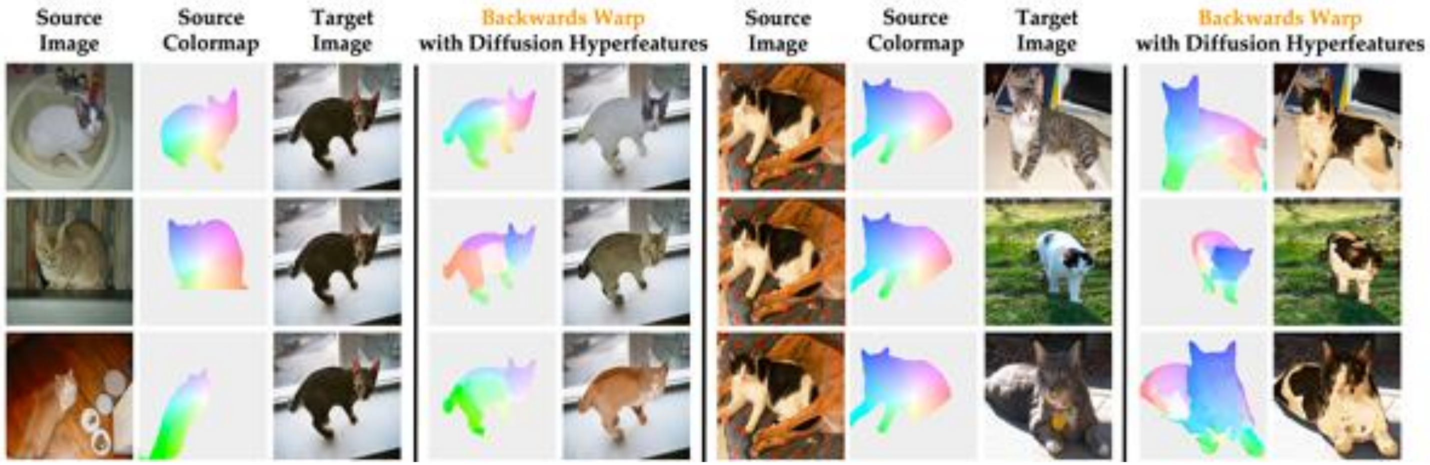
Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

Method 3

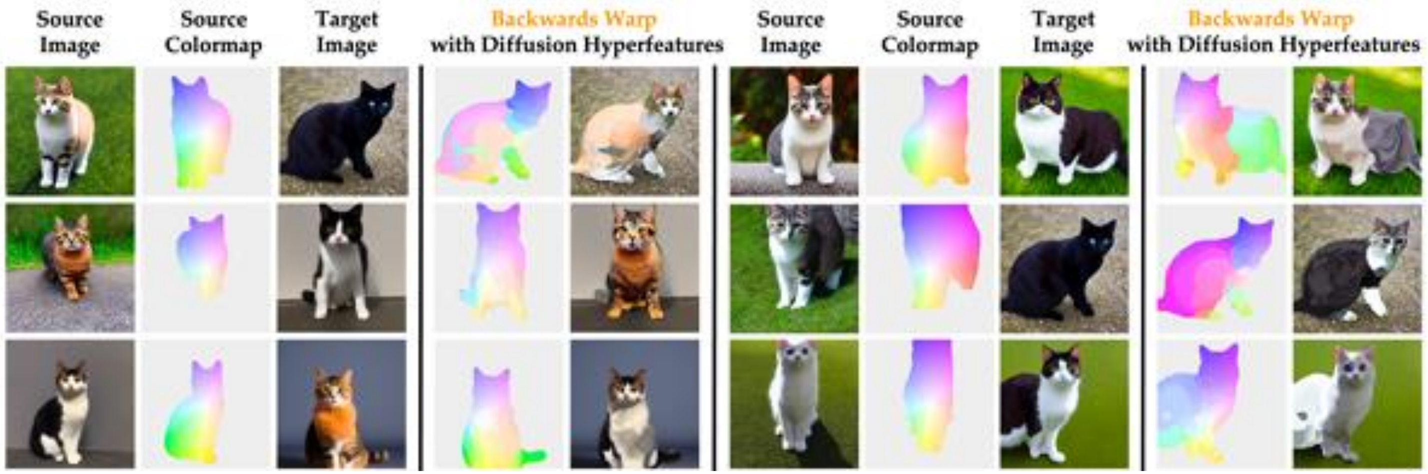
Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ **Experiments**

2) Dense Warping



Real images



Synthetic images

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

Method 3

Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence

❖ Experiments

3) Ablation – Interpretable mixing weights & Model variant

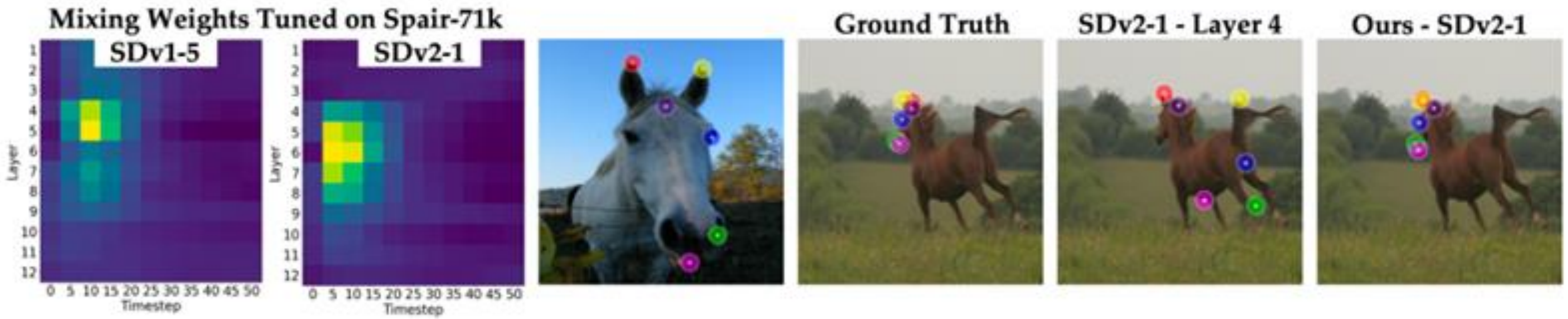


Figure 5: The learned mixing weights when aggregating SDv1-5 vs. SDv2-1 features across multiple layers and timesteps. Bright yellow denotes a high weighting, and dark blue denotes a low weighting. We also depict predicted correspondences from SDv2-1-Layer-4 vs. Ours-SDv2-1. While Layer 4 features from SDv1-5 perform well in semantic correspondence, this same layer in SDv2-1 performs extremely poorly. Our method automatically learns the best layers depending on the model variant.

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023, December). Diffusion hyperfeatures: searching through time and space for semantic correspondence. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 47500-47510).

Conclusion

❖ **Emergent Correspondence from Image Diffusion(NeurIPS, 2023)**

- Stable Diffusion의 어텐션 맵을 통해 특정 위치에 대해 최적화된 prompt embedding을 활용하여 correspondence task 수행
  - ✓ Stable Diffusion의 cross-attention layer가 correspondence estimation을 위해 사용될 수 있음을 보임

❖ **Unsupervised Semantic Correspondence Using Stable Diffusion(NeurIPS, 2023)**

- 사전 학습된 Diffusion Models 통해 추출한 Input image의 image features(DIFT)을 활용하여 correspondence task 수행
  - ✓ 이미지 픽셀 간의 cosine similarity 활용

❖ **Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence(NeurIPS, 2023)**

- Stable Diffusion 통해 추출한 모든 feature maps을 통합하는 Aggregation Network을 활용하여 correspondence task 수행
  - ✓ Diffusion Process 동안 생성되는 feature maps을 통합할 수 있는 구조 제안

고맙습니다

Appendix

Unsupervised Semantic Correspondence Using Stable Diffusion

❖ Method

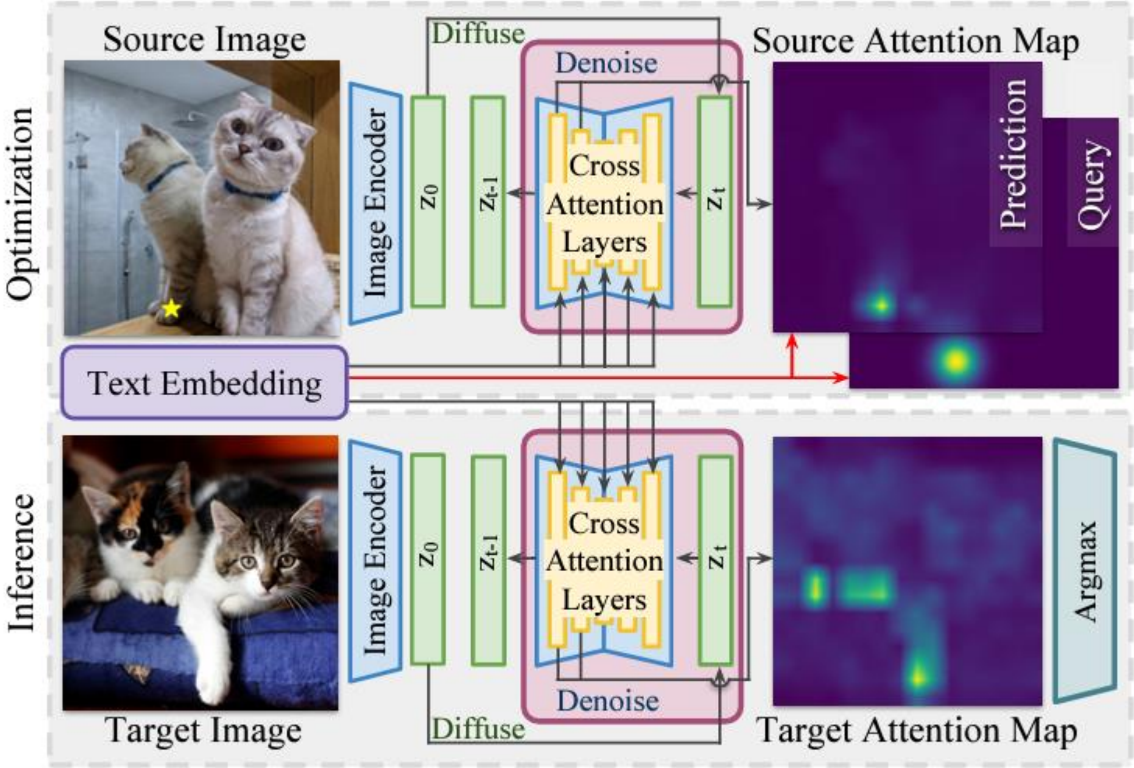
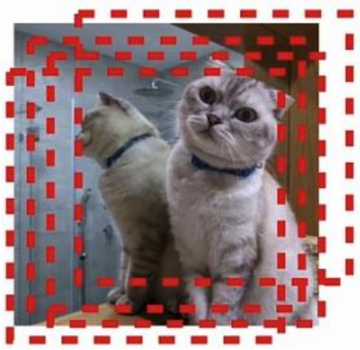


Figure 3: **Method** – (Top) Given a source image and a query point, we *optimize* the embeddings so that the attention map for the denoising step at time t highlights the query location in the source image. (Bottom) During inference, we input the target image and reuse the embeddings for the same denoising step t, determining the corresponding point in the target image as the argmax of the attention map. The architecture mapping images to attention maps is a pre-trained Stable Diffusion model [30] which is kept frozen throughout the entire process.

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

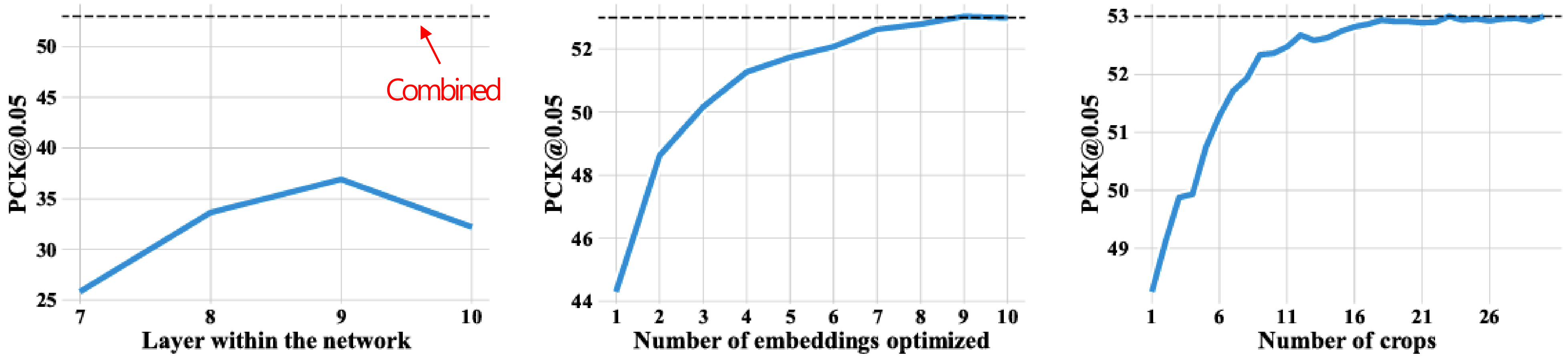
Unsupervised Semantic Correspondence Using Stable Diffusion



Random crop
- overfitting 방지

❖ Experiments

- Ablations
  - ✓ 개별 layer 성능은 매우 저조하며, 여러 layers를 동시에 사용하는 것이 도움됨을 알 수 있음
  - ✓ 많은 수의 embeddings과 crops을 사용하는 것이 성능 향상에 도움을 줌



(a) Individual layer performance (b) # embeddings vs performance (c) #crops vs performance

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., & Yi, K. M. (2023). Unsupervised Semantic Correspondence Using Stable Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

Emergent Correspondence from Image Diffusion

❖ Experiments

1) Semantic Correspondence

✓ PCK(α_{bbox} = 0.1) per image on Spair-71k

	Sup.	Method	SPair-71K Category																		
			Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	All
Fully supervised	(a)	CATs [14]	52.0	34.7	72.2	34.3	49.9	57.5	43.6	66.5	24.4	63.2	56.5	52.0	42.6	41.7	43.0	33.6	72.6	58.0	49.9
		MMNet [100]	55.9	37.0	65.0	35.4	50.0	63.9	45.7	62.8	**28.7**	65.0	54.7	51.6	38.5	34.6	41.7	36.3	77.7	62.5	50.4
		TransforMatcher [42]	**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
		SCorrSAN [35]	57.1	**40.3**	**78.3**	38.1	51.8	57.8	47.1	**67.9**	25.2	**71.3**	**63.9**	49.3	**45.3**	**49.8**	**48.8**	**40.3**	**77.7**	**69.7**	**55.3**
Weakly supervised	(b)	NCNet [69]	17.9	12.2	32.1	11.7	29.0	19.9	16.1	39.2	9.9	23.9	18.8	15.7	17.4	15.9	14.8	9.6	24.2	31.1	20.1
		CNNGeo [67]	23.4	16.7	40.2	14.3	36.4	27.7	26.0	32.7	12.7	27.4	22.8	13.7	20.9	21.0	17.5	10.2	30.8	34.1	20.6
		WeakAlign [68]	22.2	17.6	41.9	15.1	38.1	27.4	27.2	31.8	12.8	26.8	22.6	14.2	20.0	22.2	17.9	10.4	32.2	35.1	20.9
		A2Net [76]	22.6	18.5	42.0	16.4	37.9	30.8	26.5	35.6	13.3	29.6	24.3	16.0	21.6	22.8	20.5	13.5	31.4	36.5	22.3
		SFNet [45]	26.9	17.2	45.5	14.7	38.0	22.2	16.4	55.3	13.5	33.4	27.5	17.7	20.8	21.1	16.6	15.6	32.2	35.9	26.3
		PMD [48]	26.2	18.5	48.6	15.3	38.0	21.7	17.3	51.6	13.7	34.3	25.4	18.0	20.0	24.9	15.7	16.3	31.4	38.1	26.5
		PSCNet [38]	28.3	17.7	45.1	15.1	37.5	30.1	27.5	47.4	14.6	32.5	26.4	17.7	24.9	24.5	19.9	16.9	34.2	37.9	27.0
		PWarpC [83]	37.4	28.8	60.8	22.9	40.5	29.4	22.8	60.1	19.5	37.8	38.4	27.9	32.1	29.7	29.2	20.2	44.5	50.0	35.3
no supervision	(c)	DINO [10]	43.6	27.2	64.9	24.0	30.5	31.4	28.3	55.2	16.8	40.2	37.1	32.9	29.1	41.1	22.0	26.8	36.4	26.9	33.9
		DIFT$_{adm}$ (ours)	49.7	39.2	77.5	29.3	40.9	36.1	30.5	75.5	23.7	63.7	**52.8**	49.3	34.1	**52.3**	39.3	37.3	59.6	45.4	46.3
		OpenCLIP [36]	51.7	31.4	68.7	28.4	31.5	34.9	**36.1**	56.4	21.1	44.5	41.5	41.2	41.2	51.8	21.7	28.6	46.3	20.7	38.4
		DIFT$_{sd}$ (ours)	**61.2**	**53.2**	**79.5**	**31.2**	**45.3**	**39.8**	33.3	**77.8**	**34.7**	**70.1**	51.5	**57.2**	**50.6**	41.4	**51.9**	**46.0**	**67.6**	**59.5**	**52.9**

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

Emergent Correspondence from Image Diffusion

❖ Experiments

1) Semantic Correspondence

✓ PCK($\alpha_{bbox} = 0.1$) per point on Spair-71k

Sup.	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	Mean	All
		SPair-71K Category																			
Weakly supervised (b)	NBB [1, 26]	29.5	22.7	61.9	26.5	20.6	25.4	14.1	23.7	14.2	27.6	30.0	29.1	24.7	27.4	19.1	19.3	24.4	22.6	27.4	-
	GANgealing [62]	-	37.5	-	-	-	-	-	67.0	-	-	23.1	-	-	-	-	-	-	57.9	-	-
	NeuCongeal [58]	-	29.1	-	-	-	-	-	53.3	-	-	35.2	-	-	-	-	-	-	-	-	-
	ASIC [26]	57.9	25.2	68.1	24.7	35.4	28.4	30.9	54.8	21.6	45.0	47.2	39.9	26.2	48.8	14.5	24.5	49.0	24.6	36.9	-
no supervision (c)	DINO [10]	45.0	29.5	66.3	22.8	32.1	36.3	31.7	54.8	18.7	43.1	39.2	34.9	31.0	44.3	23.1	29.4	38.4	27.1	36.0	36.7
	DIFT$_{adm}$ (ours)	51.6	40.4	77.6	30.7	43.0	47.2	42.1	74.9	26.6	67.3	**55.8**	52.7	36.0	**55.9**	46.3	45.7	62.7	47.4	50.2	52.0
	OpenCLIP [36]	53.2	33.4	69.4	28.0	33.3	41.0	41.8	55.8	23.3	47.0	43.9	44.1	43.5	55.1	23.6	31.7	47.8	21.8	41.0	41.4
	DIFT$_{sd}$ (ours)	**63.5**	**54.5**	**80.8**	**34.5**	**46.2**	**52.7**	**48.3**	**77.7**	**39.0**	**76.0**	54.9	**61.3**	**53.3**	46.0	**57.8**	**57.1**	**71.1**	**63.4**	**57.7**	**59.5**

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

Emergent Correspondence from Image Diffusion

❖ Experiments

1) Semantic Correspondence

✓ Comparison on PF-WILLOW PCK per image (left) and CUB PCK per point (right)

	Sup.	Method	PCK@α_{bbox} $\alpha = 0.05$	PCK@α_{bbox} $\alpha = 0.10$
Fully supervised	(a)	SCNet [29]	38.6	70.4
		DHPF [56]	49.5	77.6
		PMD [48]	-	75.6
		CHM [54]	52.7	79.4
		CATs [14]	50.3	79.2
		TransforMatcher [42]	-	76.0
		SCorrSAN [35]	54.1	80.0
Weakly supervised	(b)	WarpC [82]	49.0	75.1
		PWarpC [83]	45.0	75.9
		GSF [39]	49.1	78.7
no supervision	(c)	DINO [10]	30.8	51.1
		DIFT$_{adm}$ (ours)	46.9	67.0
		OpenCLIP [36]	34.4	61.3
		DIFT$_{sd}$ (ours)	**58.1**	**81.2**

	Sup.	Method	PCK@$\alpha_{img} = 0.1$
Weakly supervised	(b)	GANgealing [62]	56.8
		NeuCongeal [58]	65.6
no supervision	(c)	DINO [10]	66.4
		DIFT$_{adm}$ (ours)	78.0
		OpenCLIP [36]	67.5
		DIFT$_{sd}$ (ours)	**83.5**

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

Emergent Correspondence from Image Diffusion

❖ Experiments

2) Geometric Correspondence

✓ Homography estimation accuracy [%] at 1, 3, 5 pixels on HPatches

Method	Geometric Supervision	All $\epsilon=1$	All $\epsilon=3$	All $\epsilon=5$	Viewpoint Change $\epsilon=1$	Viewpoint Change $\epsilon=3$	Viewpoint Change $\epsilon=5$	Illumination Change $\epsilon=1$	Illumination Change $\epsilon=3$	Illumination Change $\epsilon=5$
SIFT [51]	None	40.2	68.0	79.3	26.8	55.4	72.1	54.6	81.5	86.9
LF-Net [59]	Strong	34.4	62.2	73.7	16.8	43.9	60.7	53.5	81.9	87.7
SuperPoint [16]	Strong	36.4	72.7	82.6	22.1	56.1	68.2	51.9	90.8	**98.1**
D2-Net [19]	Strong	16.7	61.0	75.9	3.7	38.0	56.6	30.2	84.9	95.8
DISK [86]	Strong	40.2	70.6	81.5	23.2	51.4	67.9	58.5	91.2	96.2
ContextDesc [52]	Strong	40.9	73.0	82.2	29.6	60.7	72.5	53.1	86.2	92.7
R2D2 [66]	Strong	40.0	74.4	84.3	26.4	60.4	73.9	54.6	89.6	95.4
w/ SuperPoint kp.										
CAPS [91]	Weak	44.8	**76.3**	**85.2**	**35.7**	**62.9**	**74.3**	54.6	90.8	96.9
DINO [10]	None	38.9	70.0	81.7	21.4	50.7	67.1	57.7	90.8	97.3
DIFT_{adm} (ours)	None	43.7	73.1	84.8	26.4	57.5	**74.3**	**62.3**	90.0	96.2
OpenCLIP [36]	None	33.3	67.2	78.0	18.6	45.0	59.6	49.2	91.2	97.7
DIFT_{sd} (ours)	None	**45.6**	73.9	83.1	30.4	56.8	69.3	61.9	**92.3**	**98.1**

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

Emergent Correspondence from Image Diffusion

❖ Experiments

3) Temporal Correspondence

✓ Video label propagation results on DAVIS-2017 and JHMDB

(pre-)Trained on Videos	Method	Dataset	DAVIS $\mathcal{J}\&\mathcal{F}_m$	DAVIS $\mathcal{J}_m$	DAVIS $\mathcal{F}_m$	JHMDB PCK@0.1	JHMDB PCK@0.2
✗	InstDis [93]	ImageNet [15] w/o labels	66.4	63.9	68.9	58.5	80.2
	MoCo [30]		65.9	63.4	68.4	59.4	80.9
	SimCLR [12]		66.9	64.4	69.4	59.0	80.8
	BYOL [25]		66.5	64.0	69.0	58.8	80.9
	SimSiam [13]		67.2	64.8	68.8	59.9	81.6
	DINO [10]		71.4	67.9	74.9	57.2	81.2
	DIFT_{adm} (ours)		**75.7**	**72.7**	**78.6**	**63.4**	**84.3**
	OpenCLIP [36]	LAION [75]	62.5	60.6	64.4	41.7	71.7
	DIFT_{sd} (ours)		70.0	67.4	72.5	61.1	81.8
✓	VINCE [24]	Kinetic [11]	65.2	62.5	67.8	58.8	80.4
	VFS [95]		68.9	66.5	71.3	60.9	80.7
	UVC [49]		60.9	59.3	62.7	58.6	79.6
	CRW [37]		67.6	64.8	70.2	58.8	80.3
	Colorization [88]		34.0	34.6	32.7	45.2	69.6
	CorrFlow [44]	OxUvA [87]	50.3	48.4	52.2	58.5	78.8
	TimeCycle [92]	VLOG [21]	48.7	46.4	50.0	57.3	78.1
	MAST [43]	YT-VOS [96]	65.5	63.3	67.6	-	-
	SFC [34]		71.2	68.3	74.0	61.9	83.0

Tang, L., Jia, M., Wang, Q., Phoo, C. P., & Hariharan, B. (2023). Emergent Correspondence from Image Diffusion. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.