

Appendix

Unleashing Diffusion Transformers for Visual Correspondence by Modulating Massive Activations

A Diffusion Transformer Architecture

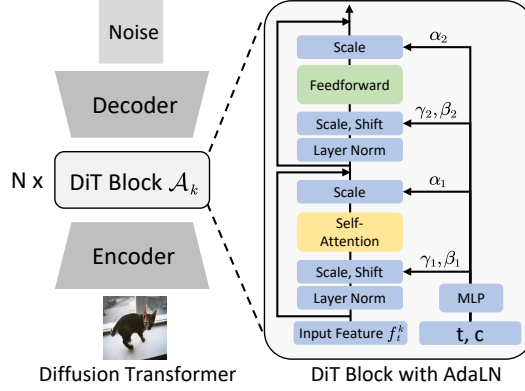


Figure 1: DiT architecture illustration.

We strictly follow the architecture used in DiT [1]. We provide an illustration of the DiT architecture as shown in Figure 1.

B More Visualization Results

B.1 Massive Activations in DiTs

To further illustrate the emergence of massive activations in DiTs, we provide additional visualization results in Figure 5. We show the LayerNorm-normalized activation magnitudes of original features from SD2-1 and various DiTs. While SD2-1 exhibits smooth activations, DiTs consistently show spikes concentrated in a few fixed dimensions across all patch tokens, revealing a fundamental difference that contributes to their degraded performance.

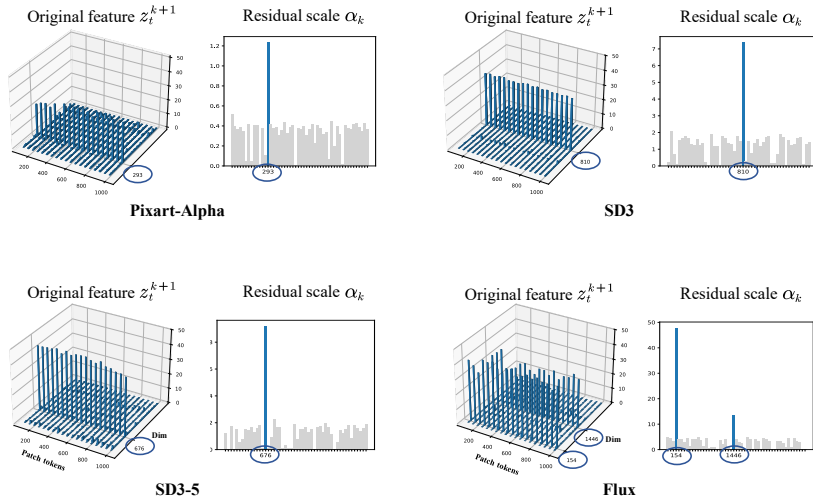


Figure 2: Massive activations dimensions align with the residual scaling factor α_k . We visualize the magnitudes for the original feature z_k^{t+1} and residual scaling factor α_k .

14 B.2 Massive Activations Dimensions Align with Residual Scaling Factor

15 In this section, we provide additional visualizations to examine the dimensional alignment between
 16 massive activations and the residual scaling factor α_k from the AdaLN layer. As illustrated in Figure 2,
 17 massive activations consistently co-occur with large values of α_k in the same dimensions across all
 18 DiTs.

19 B.3 Channel-wise Modulation with AdaLN

20 To comprehensively demonstrate the impact of the built-in AdaLN in DiTs, we provide additional
 21 visualizations comparing pre-AdaLN and post-AdaLN features across various models, including SD3-
 22 5 (Figure 3), Pixart-Alpha (Figure 6), SD3 (Figure 7), and Flux (Figure 8). These results consistently
 23 show that AdaLN accurately localizes and normalizes massive activations, while enhancing feature
 24 semantics and discrimination through effective channel-wise modulation.

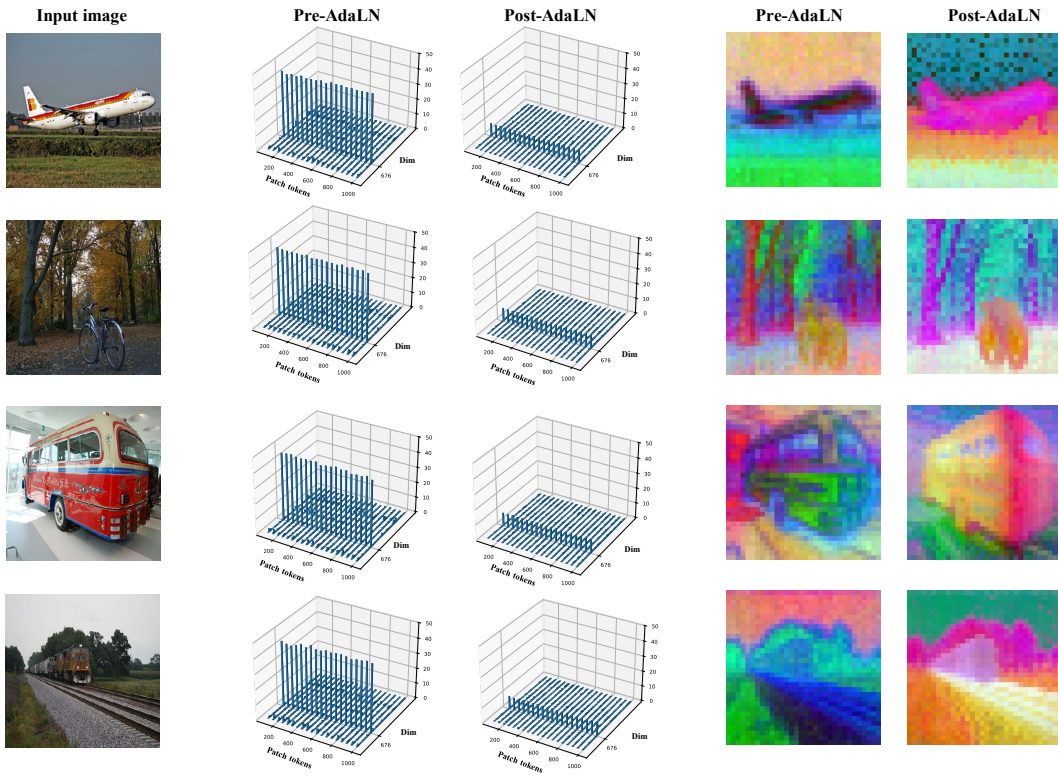


Figure 3: Comparisons of pre-AdaLN and post-AdaLN features in SD3-5.

25 C Further Implementation Details

26 **Configurations of different DiTs.** We employ the pre-trained Diffusion Transformers (DiTs) as
 27 a feature extractor for semantic correspondence. Formally, we decompose the universal feature
 28 extraction process in DiTs into two stages: (1) extracting the original feature z_t^k from the DiT block
 29 \mathcal{A}_k , and (2) modulating it via adaptive channel-wise scaling and shifting using the AdaLN layer.
 30 For Pixart-alpha [2], SD3 [3], and SD3-5 [3], we first extract the pre-AdaLN feature $z_t^{(k,2)}$ and then
 31 activate it as follows.

$$z_t^{(k,2)} = \mathcal{A}_k(z_t^k), \quad \gamma_k^2, \beta_k^2 = \text{MLP}_k(t, c) \quad (1)$$

$$\hat{z}_t^k = (1 + \gamma_k^2) \text{LayerNorm} \left(z_t^{(k,2)} \right) + \beta_k^2 \quad (2)$$

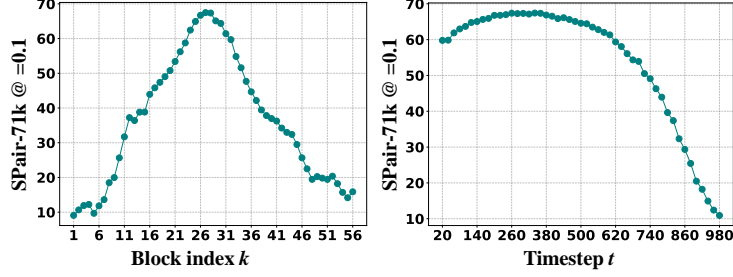


Figure 4: **Investigation of the DiT block index k and timestep t .** We report the results of DiTF_{flux} on dataset Spair-71k.

Table 1: **Configurations of different DiTs for semantic correspondence.**

Method	Layers N	Hidden size d	Timestep t	Block index k
DiTF _{pixart-α}	28	1152	141	14
DiTF _{SD3}	24	1536	340	9
DiTF _{SD3-5}	38	2432	380	23
DiTF _{flux}	57	3072	260	28

As some of the Flux [4] model’s blocks contain only one group of AdaLN-zero layer, we extract pre-AdaLN feature $z_t^{(k,1)}$ and then activate it as follows.

$$z_t^{(k,1)} = \mathcal{A}_k(z_t^k), \quad \gamma_k^1, \beta_k^1 = \text{MLP}_k(t, c) \quad (3)$$

$$\hat{z}_t^k = (1 + \gamma_k^1) \text{LayerNorm} \left(z_t^{(k,1)} \right) + \beta_k^1 \quad (4)$$

The configurations of different DiTs can be found in Table 1, where the total time step T is 1000. We set the input image size as 960x960 for DiTs and 840x840 for the model DINOv2.

Investigation on block index k and timestep t . The layer index k and the timestep t are critical hyperparameters that influence the quality of the extracted features from Diffusion Transformers (DiTs). Previous studies [5, 6] have conducted thorough investigations to identify the optimal layer index k and timestep t for Stable Diffusion. To explore the effects of varying k and t in DiTs, we conducted grid search experiments to identify the optimal hyperparameters. The results are presented in Figure 4. The figure reveals that features extracted from the middle layer achieve optimal performance across different DiTs. Furthermore, feature extraction in DiTs is robust to the timestep for semantic correspondence tasks, as a wide range of t achieves excellent performance.

Integration of DINOv2 feature. We extracted DINOv2 features from the token facet of the 11th layer of the model. We then concatenated the DiT’s features with the DINOv2 features in the channel dimension. To improve the efficiency of correspondence calculation, we computed Principal Component Analysis (PCA) across the pair of images for the features extracted from DiT, as follows:

$$\tilde{z}^s, \tilde{z}^t = \text{PCA} \left(\hat{z}^s \parallel \hat{z}^t \right) \quad (5)$$

where \hat{z}^s, \hat{z}^t are the extracted source and target image DiT features. We only apply the PCA operation to SD3-5 and Flux features due to their high dimension and set the output dimension size as 1280.

D Geometric Correspondence

To comprehensively evaluate our model DiTF, we conduct additional experiments on geometric correspondence.

Datasets. Following [5], we evaluate our model on the HPatches benchmark [26], which comprises 116 sequences: 57 with illumination changes and 59 with viewpoint variations. Adopting the approach from CAPS [14], we detect up to 1,000 key points per image and apply cv2.findHomography() to estimate homography using mutual nearest neighbor matches.

Metric. We adopt the corner correctness metric for evaluation where we compute the average error between the four estimated corners of one image and the ground-truth corners with a threshold ϵ pixels, following [14, 5].

Table 2: **Geometric correspondence results on dataset HPatches.** We report the homography estimation accuracy [%] at 1, 3, 5 pixels.

Method	Geometric Supervision	All			Viewpoint Change			Illumination Change		
		$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
SIFT [7]	None	40.2	68.0	79.3	26.8	55.4	72.1	54.6	81.5	86.9
LF-Net [8]	Strong	34.4	62.2	73.7	16.8	43.9	60.7	53.5	81.9	87.7
SuperPoint [9]		36.4	72.7	82.6	22.1	56.1	68.2	51.9	90.8	98.1
D2-Net [10]		16.7	61.0	75.9	3.7	38.0	56.6	30.2	84.9	95.8
DISK [11]		40.2	70.6	81.5	23.2	51.4	67.9	58.5	91.2	96.2
ContextDesc [12]		40.9	73.0	82.2	29.6	60.7	72.5	53.1	86.2	92.7
R2D2 [13]		40.0	74.4	84.3	26.4	60.4	73.9	54.6	89.6	95.4
<i>w/ SuperPoint kp.</i>										
CAPS [14]	Weak	44.8	76.3	85.2	35.7	62.9	74.3	54.6	90.8	96.9
DINO [15]	None	38.9	70.0	81.7	21.4	50.7	67.1	57.7	90.8	97.3
OpenCLIP [16]		33.3	67.2	78.0	18.6	45.0	59.6	49.2	91.2	97.7
DIFT [5]		45.6	73.9	83.1	30.4	56.8	69.3	61.9	92.3	98.1
DiTF _{flux} (ours)		41.9	70.7	79.5	22.0	50.8	63.4	62.5	91.3	96.2

Table 3: **Temporal correspondence results on DAVIS-2017.** We report the region-based similarity \mathcal{J} and contour-based accuracy \mathcal{F} for DAVIS. Pre-: Pre-trained on videos.

Pre-	Method	Dataset	DAVIS		
			$\mathcal{J} \& \mathcal{F}_m$	\mathcal{J}_m	\mathcal{F}_m
✓	MAST	YT-VOS [17]	65.5	63.3	67.6
	SFC [18]		71.2	68.3	74.0
✗	InstDis [17]	ImageNet [19] w/o labels	66.4	63.9	68.9
	MoCo [20]		65.9	63.4	68.4
	SimCLR [21]		66.9	64.4	69.4
	BYOL [22]		66.5	64.0	69.0
	SimSiam [23]		67.2	64.8	68.8
	DINO [24]		71.4	67.9	74.9
	OpenCLIP [16]	LAION [25]	62.5	60.6	64.4
	DIFT [5]		70.0	67.4	72.5
	DiTF _{flux} (ours)		72.2	69.2	75.1

Results. To comprehensively evaluate our models, we conducted geometric correspondence experiments on the HPatches benchmark [26], as detailed in Table 2. From the results, it can be observed that our model enables robust feature extraction for image pairs and draws precise geometric correspondence. Specifically, our model DiTF_{flux}(ours) achieve comparable performance 41.9% compared to the state-of-the-art model DIFT (SD-based). These results show that Diffusion Transformers can be employed as an effective feature extractor for geometric correspondence.

E Temporal Correspondence

In addition, we conduct experiments to verify the temporal correspondence capability of our DiTF. Specifically, we investigate DiTF’s performance on video object segmentation and pose tracking tasks, employing DiTs as a feature extractor for correspondence.

Datasets. We conduct experiments on the challenge video dataset: DAVIS-2017 video instance segmentation benchmark [27], following [5].

Metric. Following [5, 28], we adopt the region-based similarity \mathcal{J} and contour-based accuracy \mathcal{F} as the performance metric where we segment the nearest neighbors between the consecutive video frames based on the representation similarity.

Results. The temporal correspondence results can be found in Table 3. From the results, it can be observed that our model exhibits a superior capability to extract video frame features for temporal correspondence. Specifically, our model achieves 72.2% on the dataset DAVIS, which outperforms the previous state-of-the-art DIFT by 2.2%, demonstrating the superior effectiveness of our model.

79 **F Qualitative Results on AP-10K**

80 We show the qualitative comparison of our Diffusion Transformer model DiTF with both Stable
81 Diffusion (SD) and SD+DINO [6] in AP-10K intra-species (Figure 9), cross-species (Figure 10),
82 and cross-family (Figure 11) subset.

83 **G Limitations and Future Work**

84 **Lightweight Fine-Tuning for Enhanced Representation.** In this work, we propose a training-free
85 framework that leverages the built-in AdaLN in DiTs to suppress massive activations and enhance
86 feature semantics and discrimination. While our approach demonstrates strong performance without
87 additional training, further research could explore lightweight fine-tuning strategies to better adapt
88 DiTs for representation learning and fully unlock their potential as feature extractors.

89 **Understanding Massive Activations in Generation Tasks.** In this work, we identify and characterize
90 massive activations in Diffusion Transformers (DiTs) from a representation learning perspective,
91 aiming to mitigate their undesirable impact on discriminative feature extraction. Our focus is
92 orthogonal to the generative aspect of DiTs. We believe that further exploration of massive activations
93 from the viewpoint of visual generation could yield valuable insights and potentially enhance
94 generative performance.

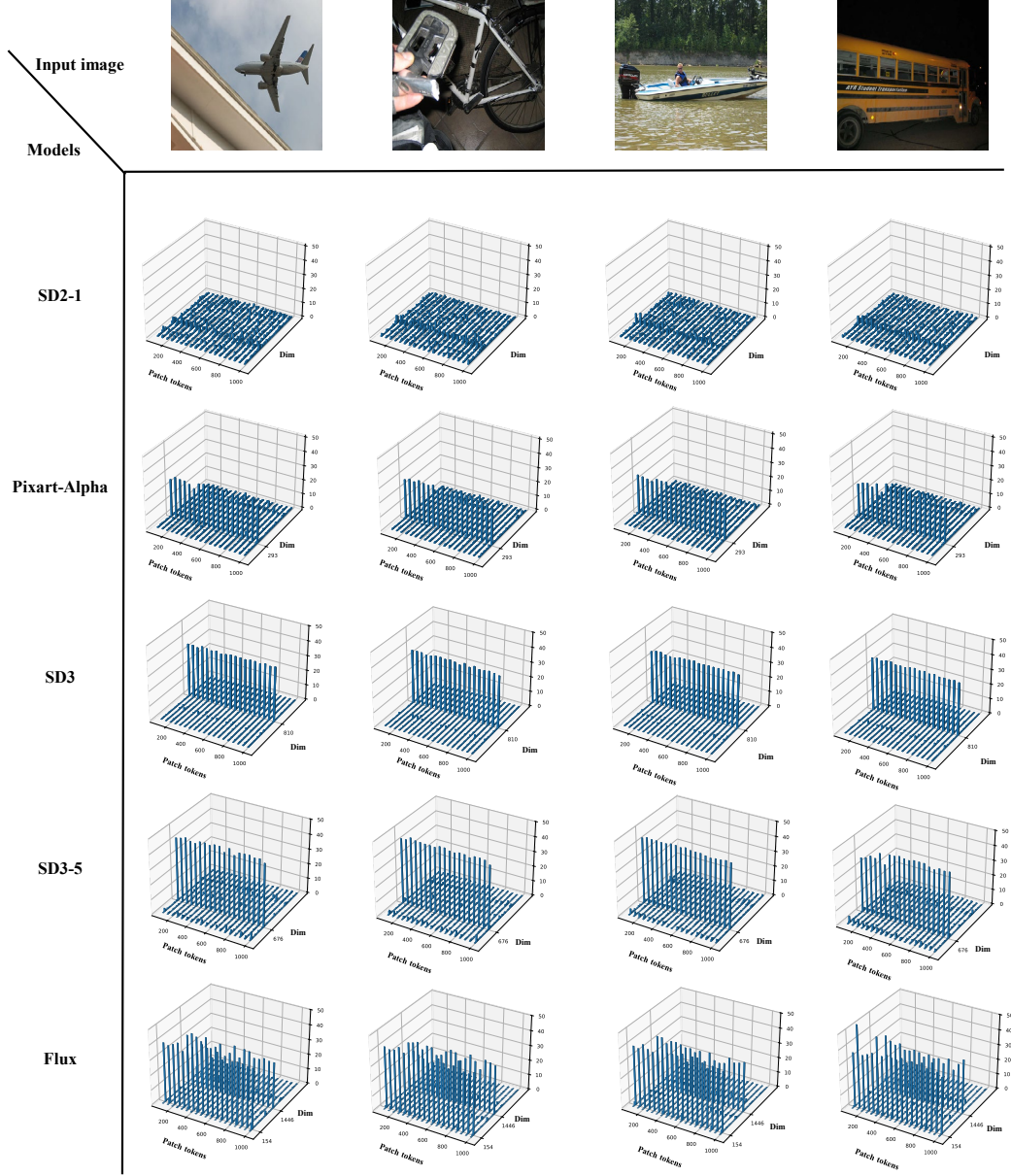


Figure 5: **Massive activations in DiTs.** SDv2-1 does not suffer from massive activations. However, all DiTs exhibit massive activations, which concentrate on very few fixed dimensions across all image patch tokens.

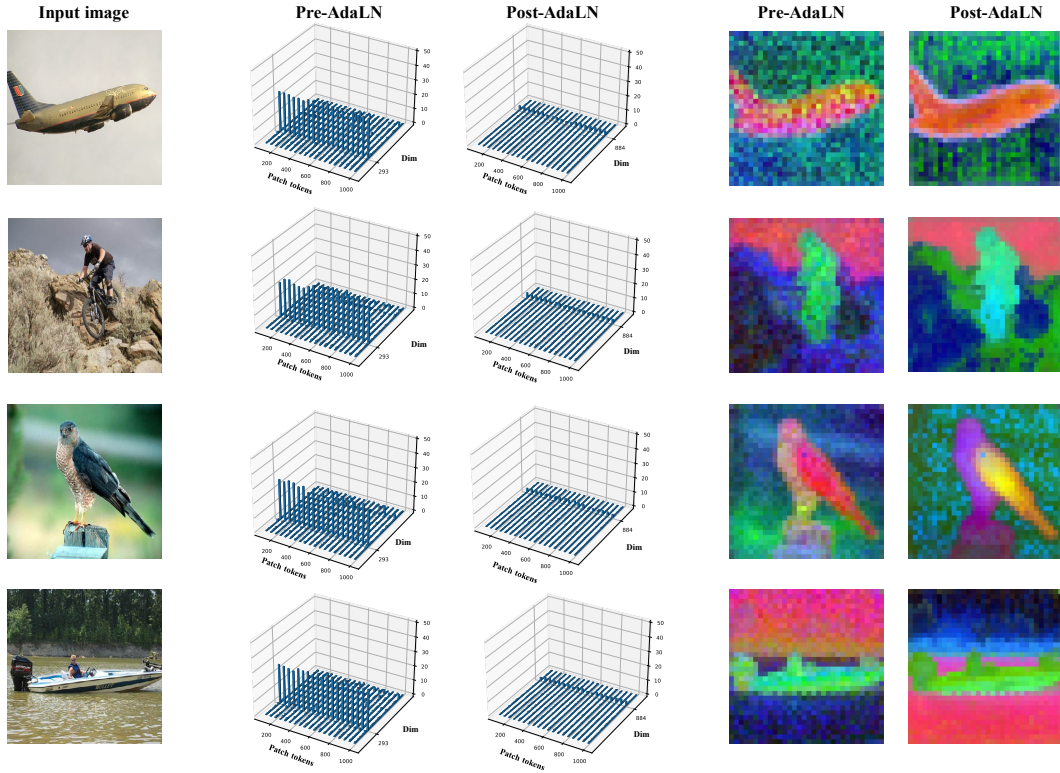


Figure 6: Comparisons of pre-AdaLN and post-AdaLN features in Pixart-Alpha.

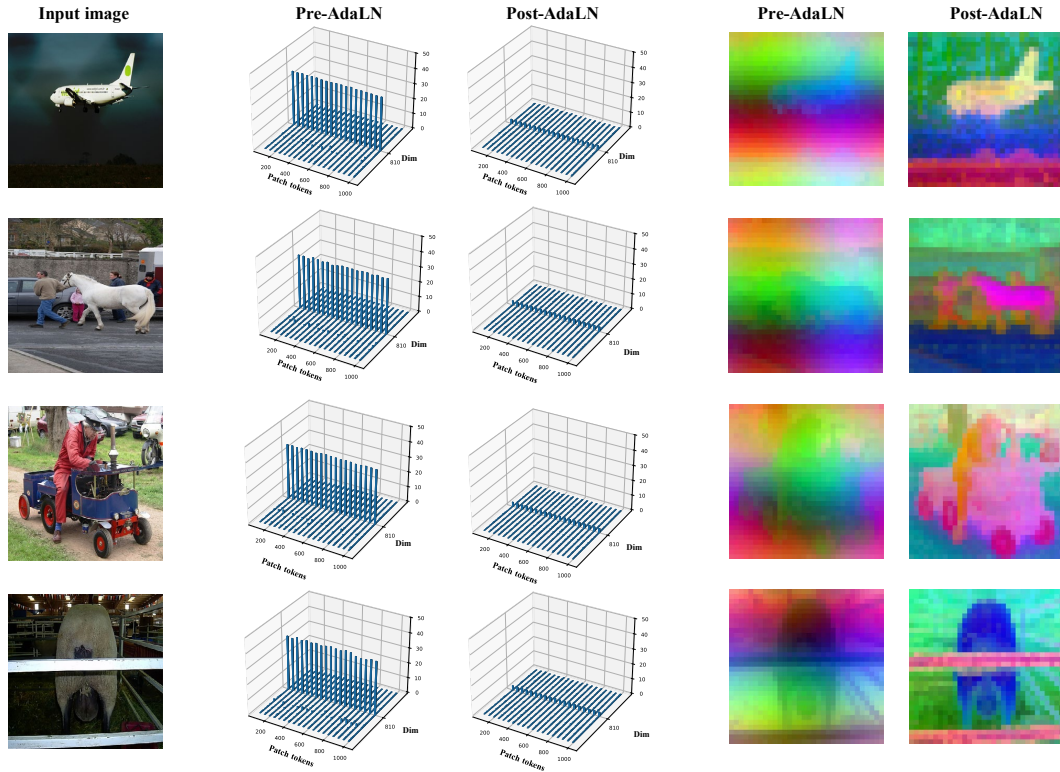


Figure 7: Comparisons of pre-AdaLN and post-AdaLN features in SD3.

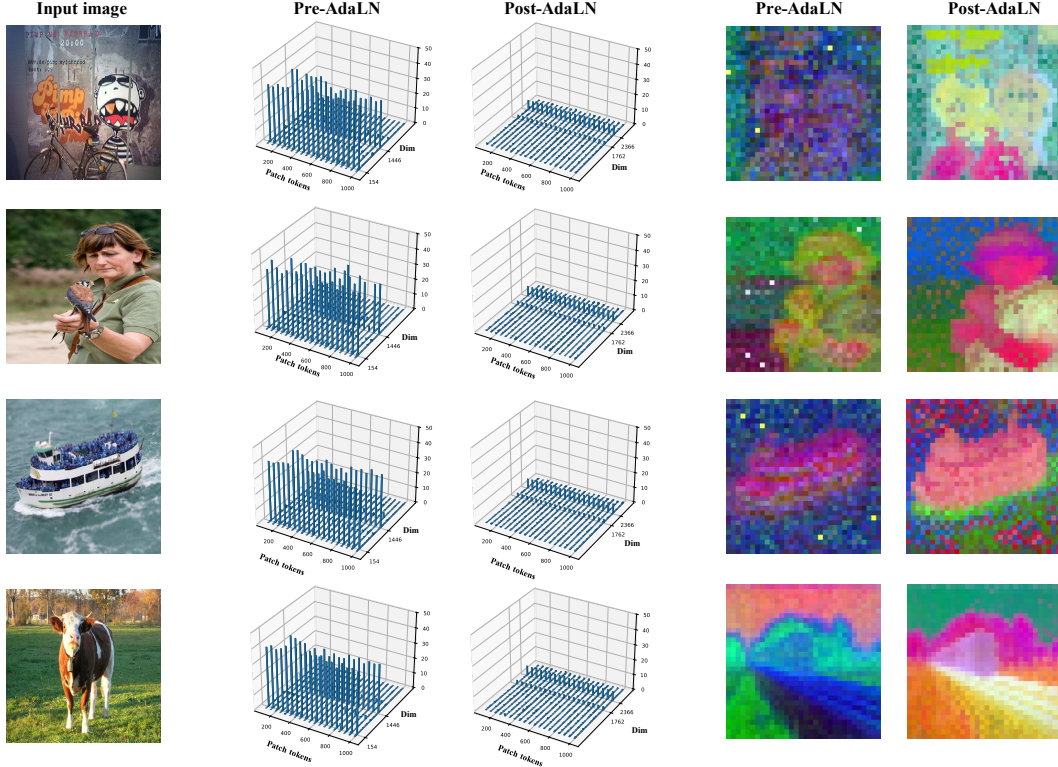


Figure 8: Comparisons of pre-AdaLN and post-AdaLN features in Flux.

References

- [1] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.
- [2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv:2310.00426*, 2023.
- [3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.
- [4] black-forest labs. Flux. <https://blackforestlabs.ai/announcing-black-forest-labs/>, 2024.
- [5] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023.
- [6] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *NeurIPS*, 2024.
- [7] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
- [8] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *NeurIPS*, 2018.
- [9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR workshops*, 2018.
- [10] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv:1905.03561*, 2019.

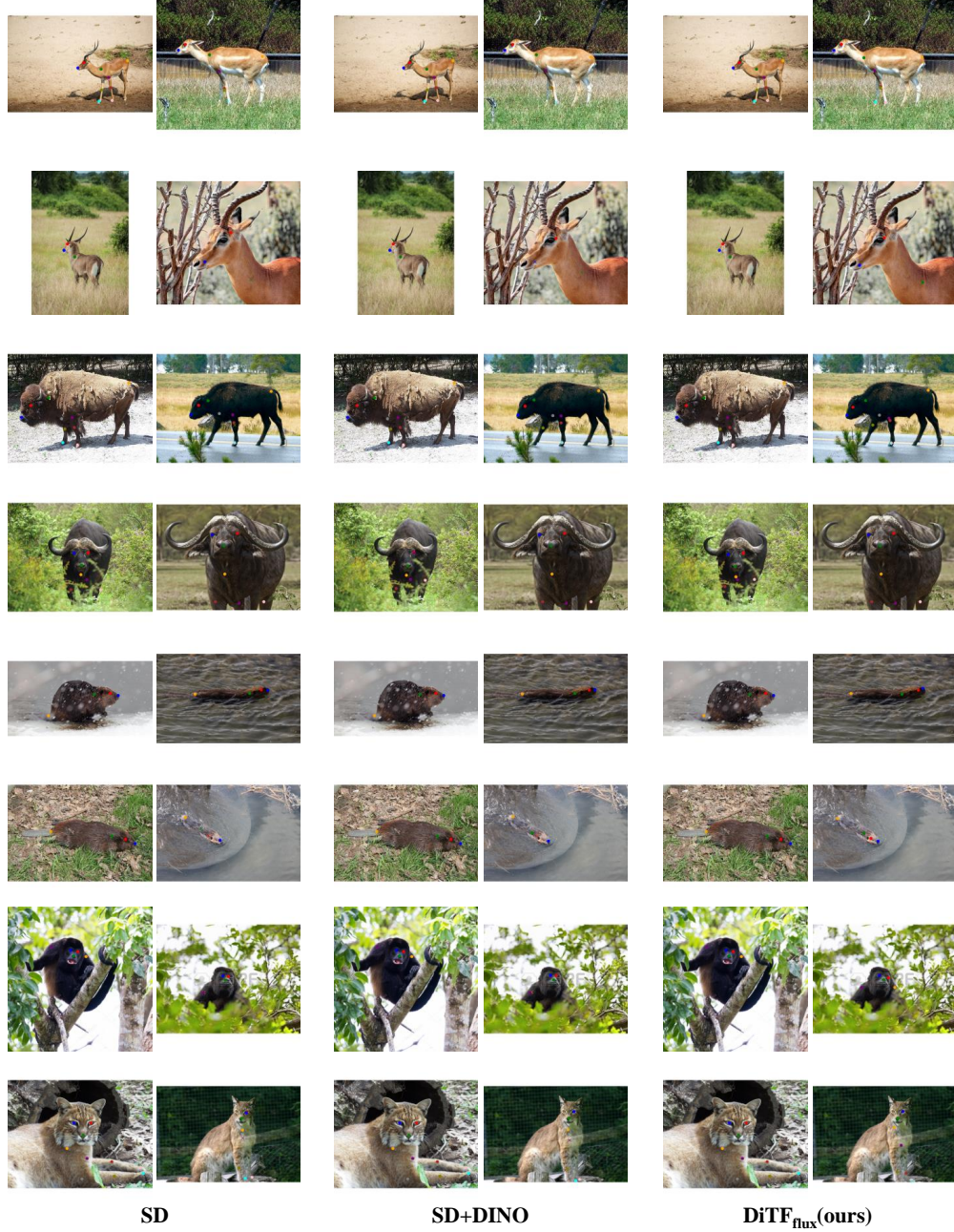


Figure 9: **Qualitative comparison on the AP-10K intra-species set.** Different colors represent different key points where circles denote correctly predicted points and crosses denote for incorrect matches.

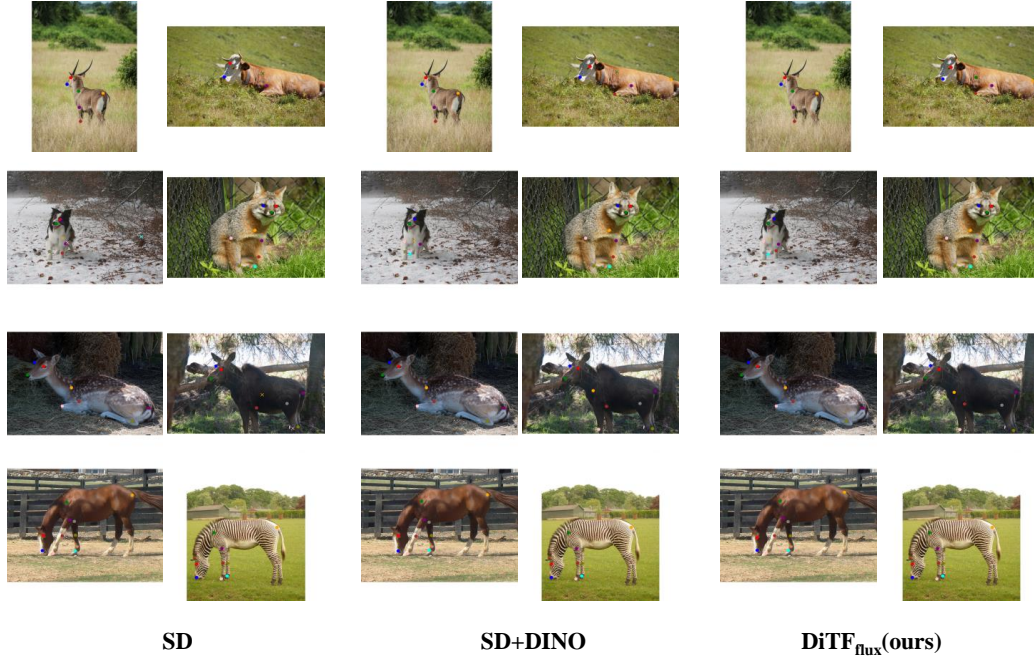


Figure 10: Qualitative comparison on the AP-10K cross-species set.

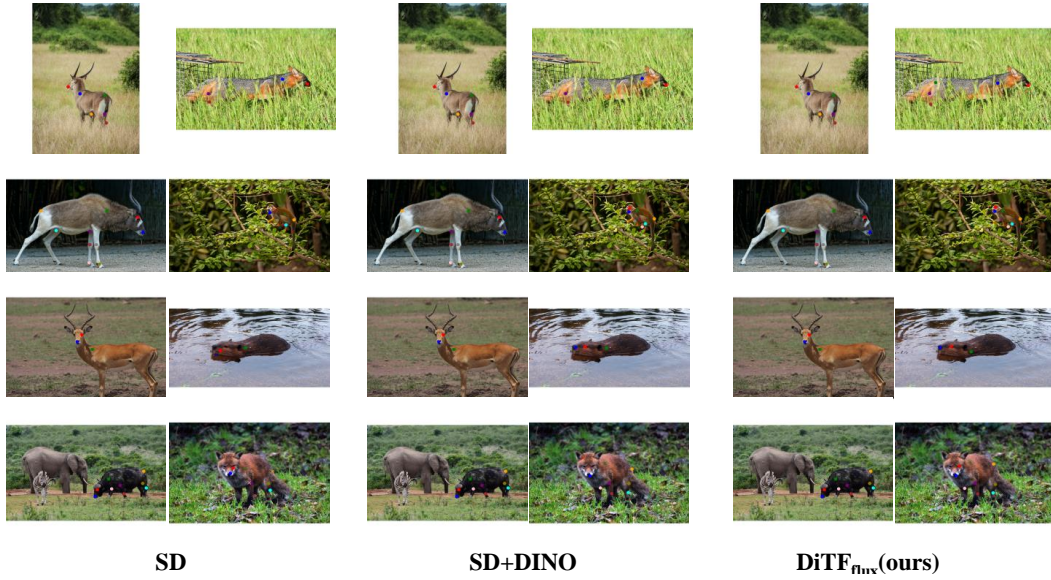


Figure 11: Qualitative comparison on the AP-10K cross-family set.

- [11] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *NeurIPS*, 2020.
- [12] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *CVPR*, 2019.
- [13] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *NeurIPS*, 2019.
- [14] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *ECCV*, 2020.
- [15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.
- [16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.
- [17] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.
- [18] Yingdong Hu, Renhao Wang, Kaifeng Zhang, and Yang Gao. Semantic-aware fine-grained correspondence. In *European Conference on Computer Vision*, pages 97–115. Springer, 2022.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [21] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- [23] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.
- [24] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
- [25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [26] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017.
- [27] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.
- [28] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems*, 32, 2019.