Supplementary Materials: Multimodal Token Fusion for Vision Transformers

Yikai Wang[1] Xinghao Chen[2] Lele Cao[1] Wenbing Huang[3] Fuchun Sun[1]✉ Yunhe Wang[2]

[1]Beijing National Research Center for Information Science and Technology (BNRist),
State Key Lab on Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University
[2]Huawei Noah's Ark Lab [3]Institute for AI Industry Research (AIR), Tsinghua University

wangyk17@mails.tsinghua.edu.cn, xinghao.chen@huawei.com, caolele@gmail.com,
hwenbing@126.com, fuchuns@tsinghua.edu.cn, yunhe.wang@huawei.com

A. Additional Results

More input modalities. In Table 7, we further evaluate our TokenFusion with more modality inputs from 1 to 4. When the number of input modalities is larger than 2, we adopt the group allocation strategy as proposed in Sec. 3.4 of our main paper. By comparison, the performance is consistently improved when using more modalities, and TokenFusion is again noticeably better than CEN [5], suggesting the ability to absorb information from more modalities.

Network sharing. As mentioned in Sec. 3.4 of our main paper, we adopt shared parameters in both Multi-head Self-Attention (MSA) and Multi-Layer Perception (MLP) for the fusion with homogeneous modalities, and rely on modality-specific Layer Normalization (LN) layers to uncouple the normalization process. Such network sharing technique is evaluated by our experiments including multimodal image-to-image translation (in Sec. 4.1) and RGB-depth semantic segmentation (in Sec. 4.2), which largely reduces the model size, and also enables the reuse of attention weights for different modalities. In Table 8, we further conduct ablation studies to demonstrate the effectiveness of our network sharing scheme. Fortunately, the comparison indicates that our default setting (*i.e.*, Shared MSA and MLP, individual LN) achieves a win-win scenario: apart from the advantage on storage efficiency, also achieves better results than using individual MSA and MLP on both tasks. Note that further sharing LN layers leads to the performance drop, especially on the image-to-image translation task. In addition, we adopt shared Positional Embeddings (PEs) by default for the fusion with homogeneous modalities, and we observe that sharing/unsharing PEs can achieve comparable performance in practice.

Combining TokenFusion with channel-wise fusion. Our TokenFusion detects uninformative tokens and replaces these tokens with projected and residual aggregated information, and utilizes these tokens for multimodal fusion. We may further combine TokenFusion with an orthogonal method by channel-wise pruning which automatically detects uninformative channels. Different from the token-wise fusion method in TokenFusion, the channel-wise fusion is not conditional on input features. Inspired by CEN [5], we leverage the scaling factors γ of layer normalization (LN) to perform channel-wise pruning, and apply sparsity constraints on γ. LN in transformers performs normalization on its input $\boldsymbol{x}_{m,l}$. To prune uninformative channels, we add a channel-wise pruning loss $\sum_{m=1}^{M}\sum_{l=1}^{L}|\boldsymbol{\gamma}_m^l|$ to the main loss in Eq. (5) (main paper). The overall loss function is

$$\mathcal{L} = \sum_{m=1}^{M}\Big(\mathcal{L}_m + \lambda_1 \sum_{l=1}^{L}\big|s^l(\boldsymbol{e}_m^l)\big| + \lambda_2 \sum_{l=1}^{L}|\boldsymbol{\gamma}_m^l|\Big), \quad (1)$$

where λ_1, λ_2 are hyper-parameters for balancing different

Modality	CEN [5]	Ours (Ti)	Ours (S)
Depth	113.91/5.68	108.16/5.50	97.13/4.97
Normal	108.20/5.42	112.25/5.77	100.29/5.02
Texture	97.51/4.82	99.70/5.14	94.92/4.38
Shade	100.96/5.17	104.73/5.43	97.35/4.77
Depth+Normal	84.33/2.70	71.82/2.36	64.20/1.69
Depth+Normal+Texture	60.90/1.56	53.17/1.22	42.54/0.93
Depth+Normal+Texture+Shade	57.19/1.33	47.69/1.01	39.15/0.81

Table 7. Results on the Taskonomy dataset for multimodal image-to-image translation (to RGB) with $1 \sim 4$ modalities.

MSA&MLP	LN	Image translation		Seg. (NYUDv2)		
		FID	KID ($\times 10^{-2}$)	Pixel Acc.	mAcc.	mIoU
Unshared	Unshared	49.73	1.06	78.3	65.6	52.9
Shared	Shared	67.45	1.82	76.7	63.8	52.0
Shared	Unshared	43.92	0.94	78.6	66.2	53.3

Table 8. Results comparison when using different network sharing schemes for image-to-image translation (Shade+Texture→RGB) on Taskonomy and RGB-depth segmentation (seg.) on NYUDv2. Lower FID or KID values indicate better performance.

✉ Corresponding author: Fuchun Sun.

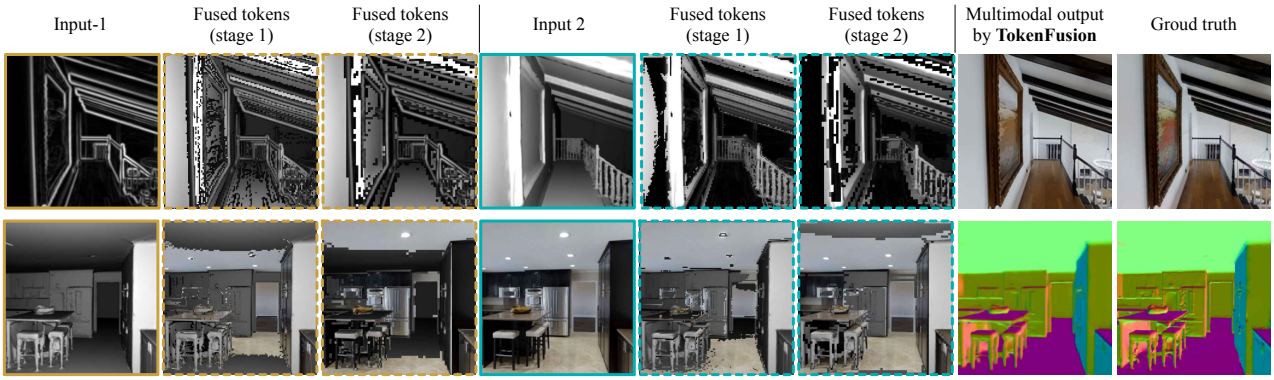

Figure 6. Additional illustrations of the token fusion process as a supplement to Fig. 4 (main paper), performed on the *validation* data split of Taskonomy. We provide two cases: Texture+Shade→RGB (first row) and Shade+RGB→Normal (second row). The resolution of all images is 256×256. We choose the last layers in the first and second transformer stages respectively. Best view in color and zoom in.

Token-wise	Channel-wise	Seg. (NYUDv2) Pixel Acc.	mAcc.	mIoU
×	×	75.2	62.5	49.7
✓	×	78.6	66.2	53.3
×	✓	77.2	65.0	52.1
✓	✓	78.8	66.6	53.8

Table 9. RGB-depth segmentation results on the NYUDv2 dataset when combining our TokenFusion with the channel-wise fusion.

Input image frames	Model	3D det. (ScanNetV2) mAP@0.25	mAP@0.5	Seconds per 100 scenes
0	Ours (L6, O256; Ti)	67.3	49.0	4.7
5	Ours (L6, O256; Ti)	67.9	50.5	5.9
10	Ours (L6, O256; Ti)	68.8	51.9	7.0

Table 10. Comparison of practical inference speed on ScanNetV2.

losses; $\boldsymbol{\gamma}_m^l$ is a vector with the length C, representing the scaling factor of LN at the l-th layer of the m-th modality.

We let $\lambda_1 = \lambda_2 = 10^{-3}$ for RGB-depth segmentation experiments. Results provided in Table 9 demonstrate that our TokenFusion can be combined with the channel-wise fusion to obtain a further improved performance. For example, the segmentation on NYUDv2 with both token-wise and channel-wise fusion achieves an additional 0.5 mIoU gain than TokenFusion. More detailed studies of such combined framework, the relation between the overall pruning rate and fusion performance gain, and the extension to fuse heterogeneous modalities are left to be the future works.

Additional visualizations. In Fig. 6, we provide another group of visualizations that depict the fused tokens under the l_1 sparsity constraints during training. We observe that fused tokens follow the regularities mentioned in our main paper, *e.g.*, the texture modality preserves its advantage at boundaries while seeking facial tokens from the shade modality.

Inference speed. In Table 10, we test the real inference speed (single V100, 256G RAM) with different numbers of input frames for 3D detection. We observe that additional time costs are mild, which is partly because the added YOLOS-Ti is a light model (with only three multi-heads).

B. More Details of Image Translation

In this part, we discuss the implementation details for our image-to-image translation task. Our implementation contains two transformers as the generator and discriminator respectively. The resolution of the generator/discriminator input or the generator prediction is 256×256. Specifically, the discriminator of our model is similar to [3], which adopts five stages with two layers for each, where the embedding dimensions and head numbers gradually double from 32 to 512 and from 1 to 16 respectively. The generator is composed of nine stages where the first five have the same configurations with the discriminator, and the last four stages have reverse configurations of its first four stages.

We adopt four kinds of evaluation metrics including Mean Square Error (MSE), Mean Absolute Error (MAE), Fréchet-Inception-Distance (FID), and Kernel-Inception-Distance (KID). Here we briefly introduce FID and KID scores. FID, proposed by [2], contrasts the statistics of generated samples against real samples. The FID fits a Gaussian distribution to the hidden activations of InceptionNet for each compared image set and then computes the Fréchet distance (also known as the Wasserstein-2 distance) between those Gaussians. Lower FID is better, corresponding to generated images more similar to the real. KID, developed by [1], is a metric similar to the FID but uses the squared Maximum-Mean-Discrepancy (MMD) between Inception representations with a polynomial kernel. Unlike FID, KID has a simple unbiased estimator, making it more reliable especially when there are much more inception features channels than image numbers. Lower KID indicates more visual similarity between real and generated images. Regarding our implementation of KID, the hidden representations are derived from the Inception-v3 [4] pool3 layer.

References

[1] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *ICLR*, 2018. 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 2

[3] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. In *NeurIPS*, 2021. 2

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2

[5] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. In *NeurIPS*, 2020. 1