

Supplementary Material: An Efficient Multi-Rater Setup Towards Personalized and Diversified Medical Image Segmentation

A. Visual results of all models

Fig. 1 and 2 compare our model’s outputs with the base models. While only D-Persona captures expert preferences (others produce minor border variations), our approach achieves superior performance with minimal annotations (one per scan) and gets better when two annotations per scan are used, evidenced by leading *GED* and *Dice* scores in personalization. More comparative personalized visualizations (provided separately) show consistent outperformance over **D-Persona Stage-II** (current SOTA) while requiring significantly fewer annotations.

B. More ablation studies

We present key ablation results on NPC-120 (all models use one annotation/scan unless marked ** for two annotations).

B.1. Personalization ablation studies

Table 1 quantifies P-Diverse Stage-I variants on NPC-120 (100 epochs). Key components:

- **E32/2xE32/E64**: One/Two 32D for different stages or 64D annotator embeddings.
- **XXX**: Modulation mode per network stages (feature extraction, encoder, and decoder) (M=modulated conv, S=SPADE, -=none).
- **F5**: 5-fold with ensemble [1].
- **3D**: 3D cascading [1].
- **D1/D2**: Small/medium discriminator (Sec. C).

Table 2 merges and extends promising variants to 250 epochs. The results favor modulated convolution (feature extraction) + SPADE (encoder/decoder).

Given the results from Tables 1 and 2, we use modulated convolution to modulate the feature extraction and SPADE to modulate the encoder and decoder.

B.2. Generalization ablation studies

We present quantitative results for stage-II trained on synthetic data from Table 2’s key experiments, including sampling effects (10/30/50) on base models and ablations.

Table 3 quantifies stage-II ablations:

- **SAP/PAP**: Setups from the key ablation studies (Sec. 5.2 from the main paper).

- **reg(P)AP**: Aligns posterior with prior during training but uses only posterior at inference. The prior prevents the latent space from collapsing.

C. Generalization loss

We augment nnU-Net’s loss with a generalization loss to encourage annotator-distinct segmentations. Experiments (D1/D2 in Table 1) use discriminators $D_\theta : x \mapsto z \in \mathbb{R}^N$ that output expert-class probabilities $p_i = e^{z_i} / \sum_j e^{z_j}$ via softmax. The cross-entropy loss $\mathcal{L} = -\log p_y$ is weighted ($\lambda = 0.01$) and added after 50 epochs.

Results show this:

- Reduces *Dice_{mean}* (undesirable for Stage-I).
- Risks training divergence.

Thus, we remove discriminators in later models and use *GED* loss instead in Stage-II. While none of the generalization loss approaches made it to the final model in stage-I, these approaches warrant future study.

D. Modulations

Fig. 3 shows the two introduced building blocks variants of the nnU-Net that capture the annotation styles. It also provides a zoomed-in view into the SPADE unit [2] from variant B.

References

- [1] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 1
- [2] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

Stage-I variant	Personalized Segmentation Performance (%)					Personalized Bounds (%)		Diversity Performance	
	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$GED \downarrow$	$Dice_{soft} \uparrow$
E32 M-	86.80	78.30	77.86	75.96	79.73	81.0	79.73	0.2071	83.17
E32 MM-	87.33	78.08	78.71	77.63	80.44	81.48	80.44	0.2017	83.75
E32 MMM	86.6	77.57	77.57	77.89	79.91	80.98	79.91	0.2007	83.59
2xE32 MMM	86.75	78.65	78.34	77.02	80.19	81.42	80.19	0.2030	83.46
E32 MS-	87.62	77.81	78.54	78.42	80.6	81.94	80.6	0.2158	83.33
E32 MSS	87.39	79.33	79.0	78.61	81.09	82.20	81.09	0.2086	83.82
2xE32 MSS	87.6	78.51	78.85	78.73	80.92	82.29	80.92	0.2094	83.85
E32 MSS D1	85.71	78.03	78.99	78.33	80.26	81.29	80.26	0.2309	83.07
E32 MSS D2	86.44	78.81	79.27	78.55	80.77	81.94	80.77	0.2169	83.56
E64 MSS	87.19	77.97	78.60	78.25	80.50	81.76	80.50	0.2121	83.24
E32 MSS F5	87.75	80.27	79.0	79.41	81.61	82.73	81.61	0.2166	83.33

Table 1. Quantitative results of the key variants in the personalization (stage-I) of our framework, P-Diverse. All of the experiments are done using only one annotation per scan. All experiments are trained for 100 epochs.

Stage-I variant	Personalized Segmentation Performance (%)					Personalized Bounds (%)		Diversity Performance	
	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$GED \downarrow$	$Dice_{soft} \uparrow$
E32 MSS F5	88.52	80.63	79.5	79.38	82.0	82.44	82.0	0.2096	84.22
2xE32 MSS F5	88.49	80.4	79.90	78.93	81.93	82.97	81.93	0.2116	84.14
E32 SSS F5	83.56	80.89	78.5	79.01	80.49	81.68	80.49	0.3025	81.57
3D E32 MSS F5	88.45	80.09	80.69	79.72	82.24	83.33	82.24	0.1922	85.13
** E32 MSS F5	91.05	80.83	80.13	80.30	83.07	84.10	83.07	0.1825	85.76

Table 2. Quantitative results of the key variants in the personalization (stage-I) of our framework, P-Diverse. All of the experiments are done using only one annotation per scan unless the experiments denoted by **. All experiments are trained for 250 epochs.

Stage-II variant	$GED \downarrow$	$Dice_{soft} \uparrow$	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$
+ P-Diverse (Stage II, #10) (SAP)	0.209	84.28	81.4	81.33
+ P-Diverse (Stage II, #30) (SAP)	0.1961	84.96	81.93	81.93
+ P-Diverse (Stage II, #50) (SAP)	0.1901	84.92	82.27	82.26
+ P-Diverse (Stage II, #10) (PAP)	0.2284	83.70	81.0	81.0
+ P-Diverse (Stage II, #30) (PAP)	0.2072	84.35	81.84	81.84
+ P-Diverse (Stage II, #50) (PAP)	0.1928	84.88	82.36	82.36
+ P-Diverse (Stage II, #10) (reg(P)AP)	0.212	83.87	80.95	80.88
+ P-Diverse (Stage II, #30) (reg(P)AP)	0.1937	84.85	81.99	81.97
+ P-Diverse (Stage II, #50) (reg(P)AP)	0.2054	83.94	81.95	81.95
+ [3D E32 MSS F5] \rightarrow P-Diverse (Stage II, #10) (PAP)	0.242	83.32	80.43	80.39
+ [3D E32 MSS F5] \rightarrow P-Diverse (Stage II, #30) (PAP)	0.2314	83.25	81.28	81.27
+ [3D E32 MSS F5] \rightarrow P-Diverse (Stage II, #50) (PAP)	0.2223	83.87	81.67	81.67
+ P-Diverse (Stage II, #10) (SAP)**	0.1951	84.36	81.1	80.95
+ P-Diverse (Stage II, #30) (SAP)**	0.174	85.05	82.21	82.2
+ P-Diverse (Stage II, #50) (SAP)**	0.1745	84.96	82.37	82.37

Table 3. Quantitative results of the key generalization variants of the base models and our P-Diverse framework (stage-II). Our model variants are trained using the data that is generated by “E32 MSS F5” of 250 epochs variant from stage-I, except where “3D E32 MSS F5” of epochs is indicated. All of the experiments are done using only one annotation per scan unless the experiments denoted by **. All experiments are trained for 200 epochs.

Method	NPC-120					QUBIQ2021		
	Scan	Expert #1	Expert #2	Expert #3	Expert #4	Scan	Expert #1	Expert #2
U-Net (A1)								
U-Net (A2)								
U-Net (A3)						NA	NA	NA
U-Net (A4)						NA	NA	NA
CM Global								
CM Pixel								
PIONONO								
D-Persona stage-II								
P-Diverse stage-I* (Ours)								
P-Diverse stage-I** (Ours)						NA	NA	NA

Figure 1. Personalization segmentations visualizations. The yellow (light) segment represents the expert label, and the overlapping orange (dark) segment represents the predicted personalized segment from the model trained on only one annotation per scan. * and ** refers to the model trained with one and two annotations per scan respectively, or all annotation otherwise.

