SUPPLEMENTARY FOR FINE-GRAINED FEW-SHOT RECOGNITION BY DEEP OBJECT PARSING

Anonymous authors
Paper under double-blind review

APPENDIX

A DERIVATION OF PARSE

As mentioned in Sec. 3 of the main paper, estimating part expression and location leads to two coupled optimization problems.

$$z_p(\mu) = \arg\min_{\beta} \sum_{c \in C} \|\phi_{c,M(\mu)} - D_{p,c}\beta_c\|^2 + \lambda\|\beta\|_1. \tag{1}$$

$$\mu_p = \arg\min_{\mu \in [G]\times[G]} \left[L_p(\mu) \triangleq \sum_{c \in C} \|[\phi_c]_{M(\mu)} - D_{p,c} z_{p,c}(\mu)\|^2 + \lambda \|z_p(\mu)\|_1 \right] \tag{2}$$

For solving the above, we first approximate the solution to Eq. (1) by optimizing the reconstruction error and subsequently thresholding. As mentioned in the main paper, this is closely related to thresholding methods employed in LASSO (Hastie et al., 2001). So, first we solve

$$z'_p(\mu) = \arg\min_{\beta} \sum_{c \in C} \|\phi_{c,M(\mu)} - D_{p,c}\beta_c\|^2$$

As a reminder, the subscript $M(\mu)$ refers to the projection of ϕ_c onto the support of $M(\mu)$, which is an $s \times s$ grid centered at μ. The quadratic form of the above optimization problem, gives us an explicit solution.

$$z'_{p,c}(\mu) = \frac{(D_{p,c} * \delta_\mu) : \phi_c}{\|D_{p,c}\|^2} = \frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|^2} \tag{3}$$

where $\delta_\mu(v) = \delta(\mu - v), v \in [G] \times [G]$ is a dirac delta centered at μ, $*$ is a convolution[1] : $D_{p,c} * \delta_\mu(v) = \sum_w D_{p,c}(w - v)\delta_\mu(v)$ and ':' is the double-dot product or the sum of all elements of an element-wise/Hadamard product.

For estimating location, we substitute z'_p into Eq. (2) resulting in an upper bound for $L_p(\mu)$, which we denote as $L'_p(\mu)$.

---

[1] Note that following terminology from signal processing this is not actually a convolution but a cross-correlation. However, the way we use this term has been accepted in literature surrounding convolutional neural networks.

$$
\begin{aligned}
L_p(\mu) \le L'_p(\mu) &= \sum_{c\in[C]} \|[\phi_{c,M(\mu)} - D_{p,c} z'_{p,c}(\mu)\|^2 + \lambda \|z'_p\|_1 \\
&= \sum_{c\in[C]} \left[\|\phi_{c,M(\mu)}\|^2 - 2(\phi_{c,M(\mu)} : D_{p,c}) z'_{p,c} + \|D_{p,c} z'_{p,c}\|^2 + \lambda |z'_{p,c}|\right] \\
&\overset{(1)}{=} \sum_{c\in[C]} \Bigg[\|\phi_{c,M(\mu)}\|^2 - 2(\phi_{c,M(\mu)} : D_{p,c}) \frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|^2} \\
&\qquad + \|D_{p,c}\|^2 . \frac{(D_{p,c} * \phi_c)(\mu)^2}{\|D_{p,c}\|^4} + \lambda \frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|^2}\Bigg] \\
&\overset{(2)}{=} \sum_{c\in[C]} \left[\|\phi_{c,M(\mu)}\|^2 - \frac{(D_{p,c} * \phi_c)(\mu)^2}{\|D_{p,c}\|^2} + \lambda \frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|^2}\right] \\
&= \sum_{c\in[C]} \Bigg[\|\phi_{c,M(\mu)}\|^2 - \frac{(D_{p,c} * \phi_c)(\mu)^2}{\|D_{p,c}\|^2} + \lambda \frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|^2} \\
&\qquad - \frac{\lambda^2}{4\|D_{p,c}\|^2} + \frac{\lambda^2}{4\|D_{p,c}\|^2}\Bigg] \qquad (4)
\end{aligned}
$$

For step (1) above, we substitute $z'_{p,c}$ from Eq. (3). For step (2), note that $D_{p,c} : \phi_{c,M(\mu)} = (D_{p,c} * \phi)(\mu)$, since $M(\mu)$ is an $s \times s$ attention map centered at μ.

From Eq. (4), by ignoring the first and the last terms and contracting the binomial squares, we get the following as our estimate for μ_p. Note that the last term is ignored because it does not depend on μ. Also, the first term $\sum_{c\in[C]} \|\phi_{c,M(\mu)}\|^2$, which is the energy across all channels varies little for different values of μ.

$$
\begin{aligned}
\mu_p = \underset{\mu\in[G]\times[G]}{\arg\min}\, L'_p(\mu) &= \underset{\mu\in[G]\times[G]}{\arg\min} - \sum_{c\in C} \left[\frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|} - \frac{\lambda}{\|D_{p,c}\|}\right]^2 \\
&= \underset{\mu\in[G]\times[G]}{\arg\max} \sum_{c\in C} ((\theta_{p,c} * \phi_c)(\mu) - \lambda_c)^2 \qquad (5)
\end{aligned}
$$

$\theta_{p,c} = D_{p,c}/\|D_{p,c}\|$, and $\lambda_c = \lambda/2\|D_{p,c}\|$ becomes a channel dependent constant. The location estimate in Eq. (5), is thus, in the form of template matching per channel.

Differentiable Estimates. As mentioned in the main paper, the above estimate (Eq. (5)) for μ_p does not provide any gradients for the parameters in $\theta_{p,c}$ or those involved in computing ϕ_c. We make the estimate differentiable in its parameters by approximating the argmax as the expectation of a softmax distribution ν_p over $[G] \times [G]$ with a low temperature T.

$$
\nu_p(\mu) = \text{softmax}\left(\frac{1}{T} \sum_{c\in C} ((\theta_{p,c} * \phi_c)(\mu) - \lambda_c)^2\right); \quad \mu_p = \mathbb{E}_{\mu\sim\nu_p} \mu \qquad (6)
$$

Substituting back the estimate of μ_p into Eq. (3) again makes z_p unusable to get gradients (since μ_p is an index in a non-continuous domain $[G] \times [G]$). One workaround is estimating z_p as an expectation over ν_p of Eq. (3) (similar to how μ_p is estimated).

$$
z'_{p,c} = \mathbb{E}_{\mu\sim\nu_p} \left[\frac{(D_{p,c} * \phi_c)(\mu)}{\|D_{p,c}\|^2}\right]; \;\; z_{p,c} = S_\zeta(z'_{p,c}) \qquad (7)
$$

However, we found a different estimate turns out to be more accurate and performs better in practice. Using the first expression from Eq. (3)

$$
z'_{p,c} = \frac{(D_{p,c} * \delta_{\mu_p}) : \phi_c}{\|D_{p,c}\|^2} \approx \frac{(D_{p,c} * \widehat{\delta}_{\mu_p}) : \phi_c}{\|D_{p,c}\|^2} \qquad (8)
$$

We make this estimate of $z_{p,c}$ differentiable by using a differentiable approximation of δ_{μ_p}, $\widehat{\delta}_{\mu_p}$ which is a low-radius ($\sigma^2 = 0.25$) gaussian centered at μ_p. With the DOP model with 1 part, this estimate (Eq. (8)) achieves an accuracy of 90.56% on 5-way 5-shot classification on the CUB dataset, while the estimate from Eq. (7) achieves an accuracy of 89.46% on the same task.

B MORE ON COMPARED METHODS

We compare DOP to state-of-the-art few-shot learning methods, including RENet (Kang et al., 2021), FRN (Wertheimer et al., 2021),TDM (Lee et al., 2022) and DeepEMD(Zhang et al., 2020) and also to methods like FOT (Wang et al., 2021), VFD (Xu et al., 2021), DN4 (Li et al., 2019) and TDM (Lee et al., 2022), which are dedicated to the fine-grained setting. To highlight the contribution of DOP , we tabulate in Tab. 1 the differences of the model design compared to prior works (Tokmakov et al., 2019; Hao et al., 2019; Zhang et al., 2020; Wu et al., 2021) in few-shot learning that also use part composition.

While there are prior works that learn recognition via object parts, and use instance-dependent reweighting, DOP is unique since it uses reconstruction with templates (RwT) as a criterion, uses a prior on the geometry of parts using part-locations and uses this geometry for comparing instances. See Tab. 1 for a tabulated comparison.

Note : There are some prior works where the notion of the term part is overloaded and is unrelated to our notion. Hence DeepEMD (Zhang et al., 2020) and LCR (Tokmakov et al., 2019) do not have a ✓under "Parts". LCR (Tokmakov et al., 2019) attempts to encode parts into image features. DeepEMD (Zhang et al., 2020) focuses in the image-distance metric based on an earth mover's distance between different parts. Here, parts are simply different physical locations in the image and not a compact collection of salient parts for recognition.

Again, FRN (Wertheimer et al., 2021) does not have a ✓under "RwT". It uses a reconstruction objective, but attempts to reconstruct query features from support. While this helps in determining belongingness to a class based on how well the support features reconstruct query, the method does not use templates that are shared across all image instances, reconstruction using which allows for low noise representations.

Table 1: Similarities and differences in high-level use of components by DOP and prior work. Parts: recognition using parts; RwT: Reconstruction with Templates; Geo: using geometry of parts for instance comparison, and incorporating prior on geometry.; Reweighting: instance dependent reweighting of matching scores.

Methods	Parts	RwT	Geo	Reweighting
LCR (Tokmakov et al., 2019)				
SAML (Hao et al., 2019)	✓			✓
DeepEMD (Zhang et al., 2020)				✓
FRN (Wertheimer et al., 2021)				
TPMS (Wu et al., 2021)	✓			✓
TDM (Lee et al., 2022)				✓
DOP (ours)	✓	✓	✓	✓

C FULL RESULTS ON CUB

We compare to more existing methods on CUB in Tab. 2.

D VISUALIZING TEMPLATES AND PART EXPRESSIONS

Some templates of the learned dictionary D_p are visualized in Fig. 1. Our model uses each template to reconstruct the original feature in the corresponding channel. We see diverse visual representations in different channels, implying that DOP learns diverse visual templates from the training set to express objects. Fig. 2 shows the activated templates for different objects. The model uses the same templates to express the same class.

Table 2: Few-shot accuracy in $\%$ on CUB (along with 95$\%$ confidence intervals). If not specified, the results are those reported in the original paper. *: results reported in (Xu et al., 2021). †: results are obtained by running the public implementation released by authors using ResNet18 backbone.

Methods	**Backbone**	*1-shot*	*5-shot*
ProtoNet(Snell et al., 2017)	ResNet18	71.88±0.91	87.42±0.48
MTL(Liu et al., 2018)*	ResNet12	73.31±0.92	82.29 ±0.51
Δ-encoder (Schwartz et al., 2018)	ResNet18	69.80±0.46	82.60±0.35
Baseline++ (Chen et al., 2019)	ResNet18	67.02±0.90	83.58±0.54
SimpleShot(Wang et al., 2019)	ResNet18	62.85±0.20	84.01±0.14
DN4(Li et al., 2019)†	ResNet18	70.47±0.72	84.43±0.45
MetaOptNet(Lee et al., 2019)*	ResNet12	75.15±0.46	87.09±0.30
AFHN(Li et al., 2020a)	ResNet18	70.53±1.01	83.95±0.63
BSNet(Li et al., 2020b)	ResNet18	69.61±0.92	83.24±0.60
DeepEMD(Zhang et al., 2020)	ResNet12	75.65±0.83	88.69±0.50
FOT(Wang et al., 2021)	ResNet18	72.56±0.77	87.22±0.46
VFD (Xu et al., 2021)	ResNet12	79.12±0.83	91.48±0.39
FRN(Wertheimer et al., 2021)	ResNet12	83.16	92.59
RENet(Kang et al., 2021)	ResNet12	79.49±0.44	91.11±0.24
TOAN(Huang et al., 2021)	ResNet12	67.17± 0.81	82.09±0.56
RAP(Hong et al., 2021)	ResNet18	83.59±0.18	90.77±0.10
LSANet(Yu et al., 2022)	Conv-64F	67.75	82.76
TDM(Lee et al., 2022)	ResNet12	83.36	92.80
HelixFormer(Zhang et al., 2022)	ResNet12	81.66±0.30	91.83±0.17
DOP	ResNet18	82.62±0.65	**92.61±0.38**
DOP	ResNet12	**83.39±0.82**	**93.01±0.43**

E ADDITIONAL ABLATION ANALYSIS

Instance-dependent reweighting based on goodness-of-fit. We use a parametric reweighting function α that reweights the distances between part expressions based on the how well the learned templates fit the part features (see Eq. 7 from the main paper). In Tab. 3, we show the effect of removing this reweighting, and simply using an average of all pairs of distances between the query and support. As we see, the reweighting function does help few shot classification accuracy.

Effect of using part-geometry for comparison. In Eq. 7 from the main paper, we use part geometries besides part expressions for computing distances. Tab. 3 also shows scenarios where we remove this component in the distance (equivalent to setting $\gamma = 0$). We see that using a distance between part geometries helps final few shot classification performance.

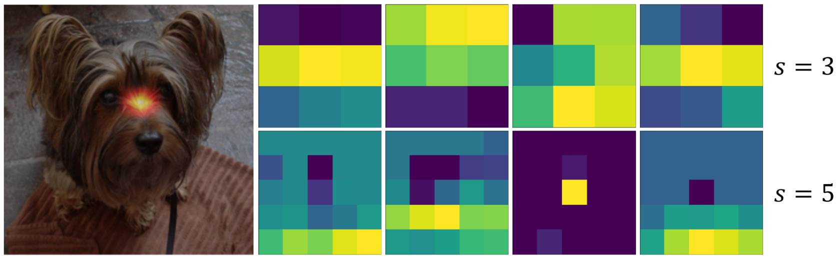

Figure 1: Exemplar templates of learned dictionary D_p. The templates shown are for randomly sampled channels for scale 3 (top) and 5 (bottom).

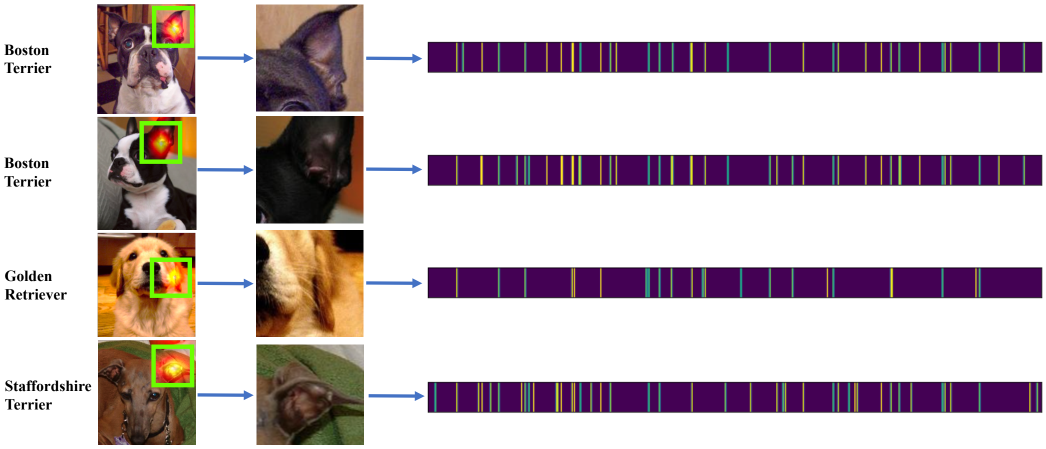

Figure 2: Template coefficients z_p of the same part for two Boston Terriers (top 2 rows), a Golden Retriever (3rd row) and a Staffordshire Terrier (4th row). Template coefficients for images of the same class are similar. Visually-similar classes (Boston Terrier and Staffordshire Terrier) share some of the same activated templates, while visually distinct classes (Golden Retriever) differ a lot on their selection of active templates.

Table 3: 5-way 5-shot accuracy on ablating components in distance computation: re-weighting function α and using part-geometry (see Eq. 7 from the main paper). Both components help FSL accuracy independently as well as together.

Part-geometry	Re-weighting	CUB	Dog	Car
		91.83	82.07	92.78
	✓	92.44	83.90	93.31
✓		91.95	83.33	93.21
✓	✓	**92.61**	**84.75**	**93.48**

REFERENCES

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao. Collect and select: Semantic alignment metric learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8460–8469, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Jie Hong, Pengfei Fang, Weihao Li, Tong Zhang, Christian Simon, Mehrtash Harandi, and Lars Petersson. Reinforced attention for few-shot learning and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 913–923, 2021.

Huaxi Huang, Junjie Zhang, Litao Yu, Jian Zhang, Qiang Wu, and Chang Xu. Toan: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):853–866, 2021.

Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8822–8833, 2021.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

SuBeen Lee, WonJun Moon, and Jae-Pil Heo. Task discrepancy maximization for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5331–5340, 2022.

Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13479, 2020a.

Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7260–7268, 2019.

Xiaoxu Li, Jijie Wu, Zhuo Sun, Zhanyu Ma, Jie Cao, and Jing-Hao Xue. Bsnet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, 30:1318–1331, 2020b.

Q Sun Y Liu, TS Chua, and B Schiele. Meta-transfer learning for few-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Rogerio Feris, Abhishek Kumar, Raja Giryes, and Alex M Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *arXiv preprint arXiv:1806.04734*, 2018.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6372–6381, 2019.

Chaofei Wang, Shiji Song, Qisen Yang, Xiang Li, and Gao Huang. Fine-grained few shot learning with foreground object transformation. *Neurocomputing*, 466:16–26, 2021.

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8012–8021, 2021.

Jiamin Wu, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8433–8442, 2021.

Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8812–8821, 2021.

Yunlong Yu, Dingyi Zhang, Sidi Wang, Zhong Ji, and Zhongfei Zhang. Local spatial alignment network for few-shot learning. *Neurocomputing*, 497:182–190, 2022.

Bo Zhang, Jiakang Yuan, Baopu Li, Tao Chen, Jiayuan Fan, and Botian Shi. Learning cross-image object semantic relation in transformer for few-shot fine-grained image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2135–2144, 2022.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12203–12213, 2020.