Supplementary Material: Mixed Patch Visible-Infrared Modality Agnostic Object Detection

Heitor R. Medeiros* David Latortue*
Eric Granger Marco Pedersoli
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
International Laboratory on Learning Systems (ILLS)
Dept. of Systems Engineering, ETS Montreal, Canada

In this supplementary material, we provide additional information to reproduce our work. This supplementary material is divided into the following sections: Detailed diagrams (Section 1), Towards the optimal ρ (Section 2), Ablation on γ (Section 3) and MiPa on different detectors (Section 4).

1. Detailed diagrams

In this section, we provide additional diagrams aimed at enhancing the comprehension of both the baselines and our method in more detail. In Figure 1, we show the simple strategy for constructing a multimodal model utilizing patches; this is our *Both* model in the main manuscript. First, the framework divides the images from both modalities (RGB and IR) into patches (yellow block). Subsequently, the extracted patches are fed into the backbone of the model (depicted in blue) and the head in pink.

In Figure 2, we present the proposed mix patches diagram. Similar to the previous diagram, we initially apply the patchify function (in yellow), followed by the mix patches function (in purple). This function receives the patches and performs a mix patches operation, such as sampling the patches from both modalities according to a uniform distribution. Finally, the backbone is illustrated in blue, and the head in pink.

Lastly, we provide an overview of an implementation of MiPa with DINO in Figure 3. While the image is similar to the previous one, we offer additional visualizations showcasing the Swin backbone alongside the modality classifier. For the sake of simplicity and to emphasize the MiPa's modality classifier and the patchify/mix patches components, we omit the detection head in the figure.

*Equal contribution. Contact: heitor.rapela-medeiros.1@ens.etsmtl.ca

2. Towards the optimal ρ

In this section, similar to the main manuscript, we provide the study of various strategies devised within this work to find the optimal approach to select the parameter ρ. This parameter represents the proportion of one modality, IR in our context, sampled during the training to facilitate optimal learning. As shown in Table 1, the variable strategy yields the most favorable results in terms of providing the optimal ρ. This effectiveness is attributed to the inherent characteristics of MiPa to act as a regularizer for the weaker modality, which is the RGB in our setup. Thus, as described, the variable strategy is the method that reached the best average across all the different APs. For example, the variable strategy was able to reach 88.5 AP_{50} in RGB, outperforming other strategies. Although its performance in IR was slightly lower than that of the Fixed strategy [$\rho = 0.25$] (achieving 97.5 AP_{50}), the variable strategy's overall mean performance was superior with 93.00 AP_{50}. This trend is similar to the other AP metrics, in which the RGB was improved, and the mean performance was better with the variable strategy.

3. Ablation on γ

In this section, we expand our comparison for different γ, in which we provide the full study on different AP metrics. The parameter γ governs the rate at which the modality invariance loss influences training. Thus, for FLIR, the best γ value was 0.05. As shown in the Table 2, we study various values of γ with steps of 0.05, selected following the GRL equation (MA module) described in our manuscript and inspired by previous works [1]. In this study, the values vary between 0.05 and 0.40, but the values may vary depending on the necessary number of epochs for training, as this function is step-dependent during training. Models that require more epochs may have larger values for γ. On FLIR, MiPa [$\gamma = 0.05$] was

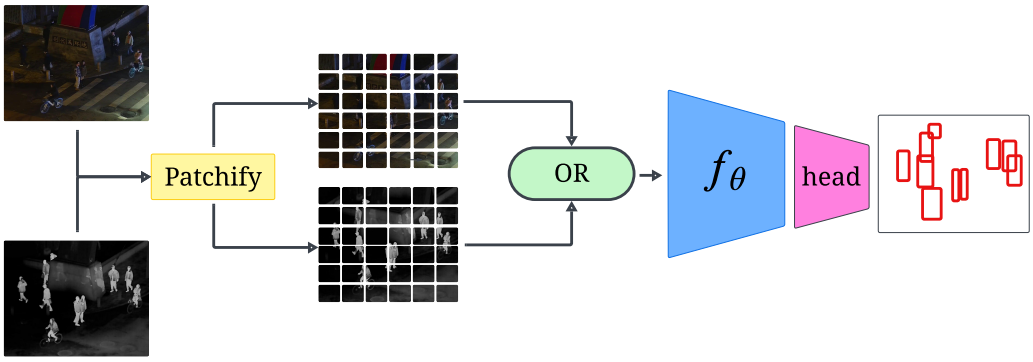

Figure 1. Our *Both* baseline for multimodal object detection learning with patches. The yellow block is the patchify function. In green, we have the block representing one or the other patch modality to use. In blue is the backbone, and in pink is the head of the detector.

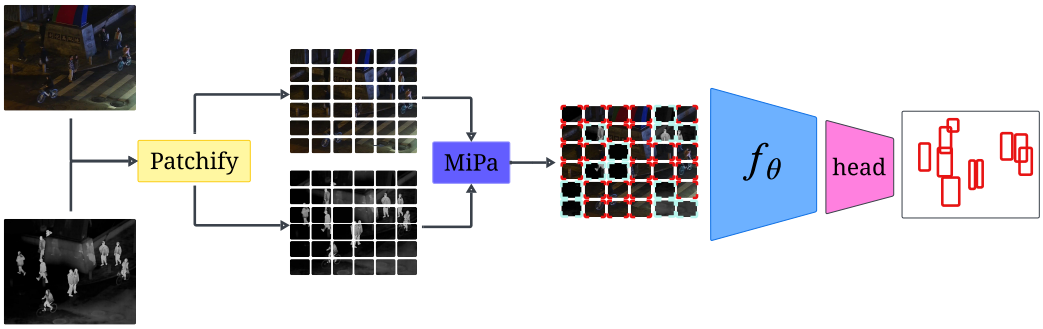

Figure 2. Mix Patches diagram: First, in yellow, is the patchify function, which is responsible for providing the patches. Second, in purple, is the mix patches function, which is responsible for mixing the patches based on a pre-defined policy, e.g., uniform distribution of both modalities. Then, in blue is the backbone, and in pink is the detection head.

Table 1. Comparison of different ratio ρ sampling methods on LLVIP. Using DINO with Swin backbone.

Model	**Dataset: LLVIP**								
	AP_{50}			**AP_{75}**			**AP**		
	RGB	**IR**	**AVG.**	**RGB**	**IR**	**AVG.**	**RGB**	**IR**	**AVG.**
Fixed [ρ=0.25]	78.9	**98.2**	88.55	41.5	78.1	59.80	42.5	66.5	54.50
Fixed [ρ=0.50]	73.0	97.6	85.30	31.1	78.1	54.60	36.0	67.0	51.50
Fixed [ρ=0.75]	77.4	97.5	87.45	40.5	76.5	58.50	42.0	65.2	53.60
Curriculum (ρ=0.25 for 4 epochs; then variable)	76.6	97.8	87.20	38.0	77.0	57.50	40.7	65.7	53.20
Curriculum (ρ=0.25 for 8 epochs; then variable)	80.1	97.8	88.95	40.9	**79.1**	60.00	43.0	**67.6**	55.30
Variable	**88.5**	97.5	**93.00**	**48.9**	77.4	**63.15**	**48.9**	66.6	**57.75**

able to outperform the other baselines with an average of 67.62 AP_{50}, which is an increase from normal MiPa with 66.52 and the best baseline with 64.72 (Both [$\rho = 0.50$]). Moreover, MiPa [$\gamma = 0.05$] reached 29.77 in terms of AP_{75}, which is an average increase from 29.25 of normal MiPa, and 27.45 from the best baseline (Both [$\rho = 0.75$]). Note that for such a case, Both [$\rho = 0.75$] was better in terms of localization (AP_{75}) in comparison with Both [$\rho = 0.50$], even though it is worse than normal MiPa and MiPa with modality agnostic layer. Finally, in terms of AP, the trend is similar, so on average, we outperform all baselines and normal MiPa, which means that we are better in terms of localization and classification in each modality simultaneously. Thus, in this section, our goal of reaching a better balance between modalities while creating a robust model is successfully achieved.

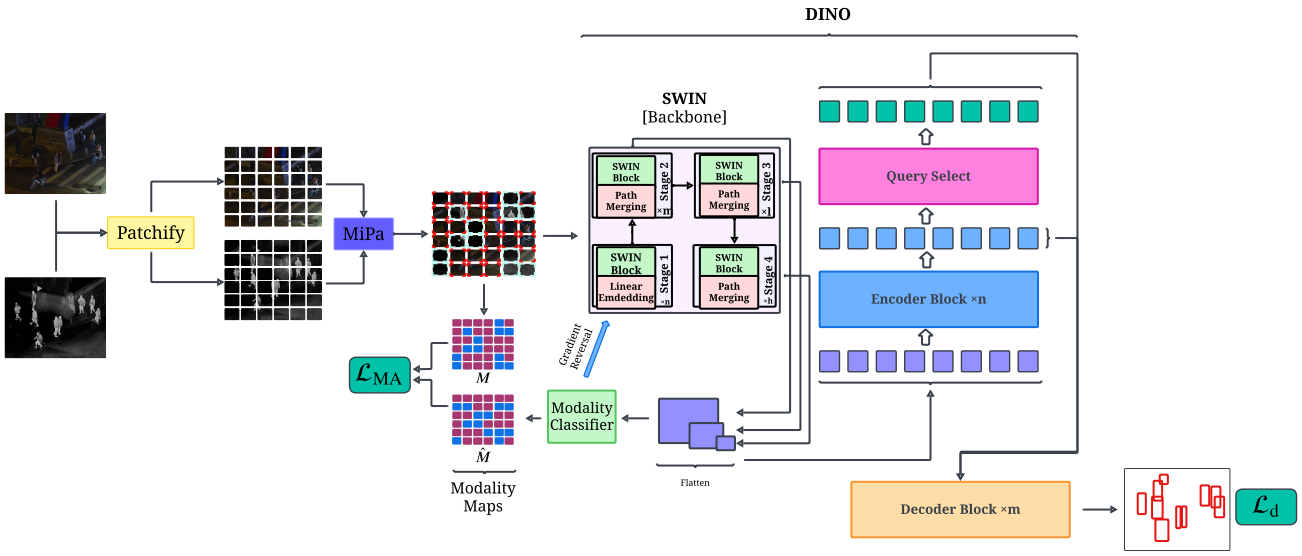

Figure 3. MiPa with DINO: First, in yellow, is the patchify function, which is responsible for providing the patches. Second, bold purple is the mixing patches function, which is responsible for mixing the patches based on a pre-defined policy, e.g., uniform distribution of both modalities. Then, we have the DINO alongside the modality classifier head for the GRL (MA module).

4. MiPa on different detectors

In this section, we present additional quantitative results, including various performance metrics measured in terms of different APs. In Table 3, we outline the results obtained using the Swin backbone for DINO and Deformable DETR across baselines, MiPa, and MiPa with a modality invariance layer. As shown, MiPa demonstrates superior performance compared to using both modalities jointly and other baselines across different datasets.

References

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation, 2015. 1

Table 2. Comparison of detection performance over different baselines and MiPa for DINO with Swin. The evaluation is done for RGB, IR, and the average of the modalities.

Model	Backbone	Modality	Test Set (Dataset: FLIR)								
			RGB			IR			Average		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
DINO	Swin	RGB	66.07 ± 0.98	27.97 ± 0.22	32.33 ± 0.47	56.60 ± 0.80	20.87 ± 0.56	26.30 ± 0.19	61.33	24.42	29.32
		IR	56.47 ± 0.79	17.00 ± 0.98	24.30 ± 0.69	70.40 ± 0.38	38.80 ± 0.66	38.97 ± 0.31	63.43	27.90	31.63
		Both [$\rho = 0.25$]	56.53 ± 0.76	18.33 ± 0.55	25.60 ± 0.33	67.57 ± 1.73	31.33 ± 2.10	34.87 ± 1.35	62.05	24.83	30.23
		Both [$\rho = 0.50$]	60.50 ± 0.66	19.60 ± 1.29	27.37 ± 0.58	68.93 ± 0.60	33.03 ± 1.32	35.90 ± 0.82	64.72	26.32	31.63
		Both [$\rho = 0.75$]	58.53 ± 0.92	19.40 ± 0.83	26.47 ± 0.75	70.43 ± 0.65	35.50 ± 1.23	37.53 ± 0.41	64.48	27.45	32.00
		MiPa	63.53 ± 1.94	22.33 ± 0.82	29.47 ± 0.92	69.50 ± 1.84	36.17 ± 0.46	37.57 ± 0.67	66.52	29.25	33.52
		MiPa [$\gamma = 0.05$]	64.80 ± 2.30	24.77 ± 1.05	30.60 ± 0.62	70.43 ± 0.53	34.77 ± 1.18	37.50 ± 0.43	**67.62**	**29.77**	**34.05**
		MiPa [$\gamma = 0.10$]	64.03 ± 2.11	24.10 ± 1.63	30.63 ± 1.22	69.63 ± 1.45	33.13 ± 1.95	36.80 ± 1.39	66.83	28.62	33.72
		MiPa [$\gamma = 0.15$]	64.27 ± 0.47	24.40 ± 0.93	30.07 ± 0.68	69.93 ± 1.02	33.83 ± 1.24	36.80 ± 0.86	67.10	29.12	33.43
		MiPa [$\gamma = 0.20$]	61.83 ± 1.39	22.83 ± 1.01	28.53 ± 0.76	69.27 ± 1.57	31.87 ± 2.02	35.73 ± 1.31	65.55	27.35	32.13
		MiPa [$\gamma = 0.30$]	62.20 ± 2.49	22.93 ± 1.35	29.10 ± 1.28	67.47 ± 2.04	32.53 ± 0.66	35.87 ± 0.69	64.83	27.73	32.48
		MiPa [$\gamma = 0.40$]	61.13 ± 2.88	22.30 ± 0.57	28.50 ± 0.99	67.93 ± 0.92	32.47 ± 0.48	35.87 ± 0.49	64.53	27.38	32.18

Table 3. Comparison of detection performance over different baselines and MiPa for DINO and Deformable DETR. The evaluation is done for RGB, IR, and the average of the modalities.

Model	Backbone	Modality	Dataset: LLVIP								
			RGB			IR			Average		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
DINO	Swin	RGB	90.87 ± 0.84	54.20 ± 1.02	51.87 ± 0.79	94.23 ± 0.57	67.13 ± 0.85	59.43 ± 0.48	92.55	60.67	55.65
		IR	66.87 ± 0.90	20.27 ± 0.98	29.03 ± 0.76	96.87 ± 0.12	73.53 ± 0.40	64.27 ± 0.12	81.87	46.90	46.65
		Both [$\rho = 0.25$]	79.73 ± 1.03	45.70 ± 0.43	44.97 ± 0.33	97.40 ± 0.22	76.03 ± 0.83	65.87 ± 0.45	88.57	60.87	55.42
		Both [$\rho = 0.50$]	82.40 ± 1.50	47.27 ± 1.65	46.43 ± 1.03	96.50 ± 0.29	74.17 ± 2.10	64.83 ± 0.96	89.45	60.72	55.63
		Both [$\rho = 0.75$]	81.23 ± 2.89	45.60 ± 2.49	45.23 ± 2.13	97.07 ± 0.25	74.73 ± 1.41	65.27 ± 0.82	89.15	60.17	55.25
		MiPa (Ours)	88.70 ± 0.45	46.67 ± 0.86	48.00 ± 0.28	96.97 ± 0.26	73.07 ± 1.42	64.30 ± 1.10	**92.83**	59.87	**56.15**
		MiPa + MA (Ours)	89.10 ± 0.28	46.60 ± 0.86	48.10 ± 0.33	96.83 ± 0.09	71.17 ± 0.70	63.17 ± 0.58	**92.97**	58.88	55.63
Def.DETR	Swin	RGB	80.00 ± 1.50	35.50 ± 0.22	40.27 ± 0.41	90.03 ± 0.87	50.37 ± 0.85	49.67 ± 0.48	85.02	42.93	**44.97**
		IR	56.10 ± 2.50	10.77 ± 1.47	21.10 ± 1.34	94.20 ± 0.08	62.20 ± 0.86	56.73 ± 0.47	75.15	36.48	38.92
		Both [$\rho = 0.25$]	51.20 ± 3.47	22.57 ± 1.96	25.70 ± 1.91	83.73 ± 16.57	54.17 ± 16.62	48.30 ± 14.93	67.47	38.37	37.00
		Both [$\rho = 0.50$]	53.57 ± 4.17	23.13 ± 2.15	26.57 ± 2.11	83.87 ± 16.17	52.67 ± 17.17	49.37 ± 12.64	68.72	37.90	37.97
		Both [$\rho = 0.75$]	53.53 ± 4.55	22.83 ± 2.72	26.5 ± 2.63	82.33 ± 18.48	51.33 ± 18.56	48.13 ± 14.03	67.93	37.08	37.32
		MiPa (Ours)	78.60 ± 0.42	23.33 ± 5.85	29.20 ± 6.37	95.20 ± 0.16	62.60 ± 0.78	56.80 ± 0.45	**86.90**	**42.97**	43.00
		MiPa + MA (Ours)	79.02 ± 0.21	24.36 ± 2.85	31.25 ± 4.32	95.36 ± 0.25	63.38 ± 0.43	57.25 ± 0.43	**87.19**	**43.87**	44.25

Model	Backbone	Modality	Dataset: FLIR								
			RGB			IR			Average		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
DINO	Swin	RGB	66.07 ± 0.98	27.97 ± 0.22	32.33 ± 0.47	56.60 ± 0.80	20.87 ± 0.56	26.30 ± 0.19	61.33	24.42	29.32
		IR	56.47 ± 0.79	17.00 ± 0.98	24.30 ± 0.69	70.40 ± 0.38	38.80 ± 0.66	38.97 ± 0.31	63.43	27.90	31.63
		Both [$\rho = 0.25$]	56.53 ± 0.76	18.33 ± 0.55	25.60 ± 0.33	67.57 ± 1.73	31.33 ± 2.10	34.87 ± 1.35	62.05	24.83	30.23
		Both [$\rho = 0.50$]	60.50 ± 0.66	19.60 ± 1.29	27.37 ± 0.58	68.93 ± 0.60	33.03 ± 1.32	35.90 ± 0.82	64.72	26.32	31.63
		Both [$\rho = 0.75$]	58.53 ± 0.92	19.40 ± 0.83	26.47 ± 0.75	70.43 ± 0.65	35.50 ± 1.23	37.53 ± 0.41	64.48	27.45	32.00
		MiPa (Ours)	63.53 ± 1.94	22.33 ± 0.82	29.47 ± 0.92	69.50 ± 1.84	36.17 ± 0.46	37.57 ± 0.67	**66.52**	**29.25**	**33.52**
		MiPa + MA (Ours)	64.80 ± 2.30	24.77 ± 1.05	30.60 ± 0.62	70.43 ± 0.53	34.77 ± 1.18	37.50 ± 0.43	**67.62**	**29.77**	**34.05**
Def.DETR	Swin	RGB	49.33 ± 1.39	13.93 ± 0.30	20.97 ± 0.53	43.77 ± 0.56	10.13 ± 0.08	17.37 ± 0.19	46.55	12.03	19.17
		IR	39.17 ± 1.48	08.57 ± 0.24	14.90 ± 0.50	59.20 ± 0.29	20.03 ± 0.33	26.93 ± 0.62	49.18	14.30	20.92
		Both [$\rho = 0.25$]	35.73 ± 4.95	08.27 ± 1.51	14.00 ± 2.38	43.00 ± 13.54	14.30 ± 5.97	19.23 ± 7.01	39.37	11.28	16.62
		Both [$\rho = 0.50$]	33.93 ± 5.15	08.23 ± 1.43	13.60 ± 2.17	43.33 ± 14.14	14.70 ± 6.34	19.63 ± 7.43	38.63	11.47	16.62
		Both [$\rho = 0.75$]	32.90 ± 3.54	07.70 ± 1.20	12.97 ± 1.65	44.13 ± 14.85	14.17 ± 6.30	19.47 ± 7.37	38.52	10.93	16.22
		MiPa (Ours)	48.00 ± 0.57	15.23 ± 0.69	20.70 ± 0.45	54.97 ± 0.90	19.80 ± 0.28	25.50 ± 0.42	**51.48**	**17.52**	**23.10**
		MiPa + MA (Ours)	48.27 ± 1.76	14.57 ± 1.05	20.63 ± 0.96	55.80 ± 0.22	21.00 ± 0.67	26.33 ± 0.39	**52.03**	**17.78**	**23.48**